ESTATÍSTICA ELEMENTAR

Este texto é uma das obras clássicas para o ensino da Estatística Metodológica. Ele não se dirige a nenhuma área específica de aplicação do método estatístico. Sua finalidade é de natureza bem mais ampla: escrito a nível elementar, destina-se a todos os cursos superiores em que se ensina, preliminarmente, a metodologia estatística e, em seguida, as suas possíveis e específicas aplicações. Nestas condições, **Estatística Elementar,** de Hoel, é uma obra básica, onde um de seus mais reconhecidos méritos é ser elaborada dentro de padrões rigorosamente didáticos.

O conjunto de temas abordados nesta obra está disposto de maneira a permitir que se estudem, preliminarmente, a natureza e os métodos estatísticos, bem como a descrição dos dados amostrais. Faz-se, também, um estudo bastante abrangente das probabilidades, com as suas várias distribuições. Desenvolvem-se, depois, explicações sobre amostragem, estimação, testes de hipóteses e correlação e regressão. Até este ponto tem-se um tratamento da matéria básica propriamente dita, isto é, dos elementos de estatística. Subseqüentemente são apresentados o que se denomina "Tópicos Adicionais", os quais servem para a ampliação do curso básico.

NOTA SOBRE O AUTOR

PAUL G. HOEL é professor de Matemática da University of California, Los Angeles.

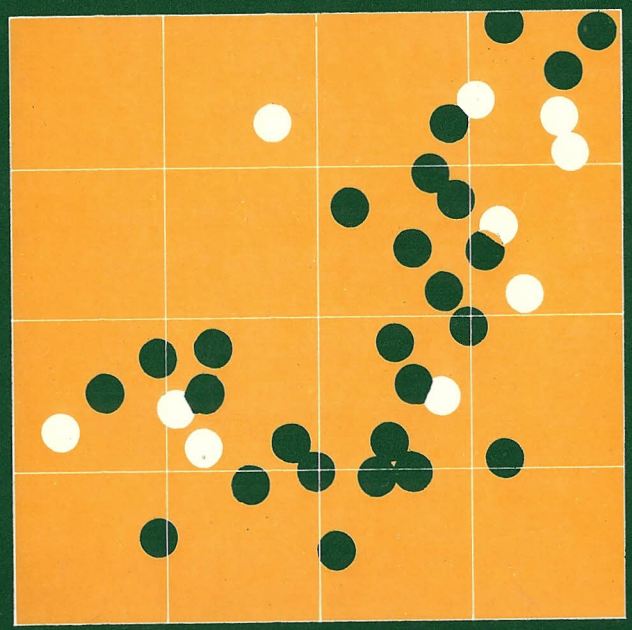
APLICAÇÃO

Livro-texto para as disciplinas ESTATÍSTICA METODOLÓGICA ou INTRODUÇÃO À ESTATÍSTICA, dos cursos de Administração, Economia, Ciências Contábeis e Ciências Sociais. Atende também à área de Ciências Exatas, como primeiro texto de referência.

publicação atlas

PAUL G. HOEL

estatística elementar



atlas

CIP-Brasil. Catalogação-na-Fonte
Câmara Brasileira do Livro, SP

H634e

Hoel, Paul Gerhard, 1905-
Estatistica elementar /por/ Paul G. Hoel; tradução de Carlos Roberto Vieira Araújo. São Paulo, Atlas, 1981

1. Estatística 2. Estatística matemática I. Título.

17. e 18. CDD-310
17. -519
18. -519.5

77-0690

Índices para catálogo sistemático:

1. Estatística 310 (17. e 18.)
2. Estatística matemática 519 (17.)
519.5 (18.)

atlas

EDITORA ATLAS S.A.
Rua Conselheiro Nébias, 1384 (Campos Elísios)
Caixa Postal 7186 — Tel.: (011) 221-9144 (PABX)
01203 São Paulo (SP)

310
HOE
EST

Paul G. Hoel

Professor de Matemática
University of California, Los Angeles

Estatística Elementar

Tradução de
CARLOS ROBERTO VIEIRA ARAÚJO
Professor de Pesquisa Operacional
Universidade Estadual de Maringá, Paraná

EDITORA ATLAS S.A.

ESTATÍSTICA ELEMENTAR

PAUL G. HOEL

Tradução autorizada da edição em língua inglesa publicada por John Wiley & Sons, Inc.

Authorized translation from English language edition published by John Wiley & Sons, Inc.

ELEMENTARY STATISTICS, 4th Edition

Copyright © 1960, 1966, 1971, 1976 by John Wiley & Sons Inc.
ALL RIGHTS RESERVED

Tradução para a língua portuguesa de
CARLOS ROBERTO VIEIRA ARAÚJO

Capa e diagramação de
PAVEL GERENCER

Copyright © 1981
EDITORA ATLAS S. A.

TODOS OS DIREITOS RESERVADOS — Nos termos da Lei que resguarda os direitos autorais, é proibida a reprodução total ou parcial, bem como a produção de apostilas a partir deste livro, de qualquer forma ou por qualquer meio — eletrônico ou mecânico, inclusive através de processos xerográficos, de fotocópia e de gravação — sem permissão, por escrito, do Editor.

1.ª Edição
8.ª tiragem — 1989

Impresso no Brasil
Printed in Brazil

SUMÁRIO

Prefácio à Quarta edição 11

A NATUREZA DOS MÉTODOS ESTATÍSTICOS, 13

1

1. Introdução, 13
2. Ilustrações, 15
3. Estimativas e Testes de Hipótese, 16
4. Probabilidade, 17
5. Organização, 18

DESCRIÇÃO DE DADOS AMOSTRAIS, 19

2

1. Introdução, 19
2. Classificação de dados, 23
3. Representação gráfica, 24
4. Descrição aritmética, 27
   4.1 — Média, 28
   4.2 — Desvio-padrão, 31
   4.3 — Interpretação do desvio-padrão, 34
5. Outras medidas descritivas, 37
6. Descrição de dados amostrais — Ilustrações adicionais, 39
7. Comentário sobre os cálculos, 44

Exercícios, 44

PROBABILIDADES, 49

3

1. Introdução, 49
2. Espaço amostral, 51
3. Probabilidade de um evento, 52
4. Probabilidade de um evento composto, 56
5. Regra da adição, 58

6. Regra da multiplicação, 61
7. Regra da multiplicação para eventos independentes, 67
8. Ilustrações, 69
9. Fórmula de Bayes, 72
10. Técnicas de contagem — Árvores, 76
11. Técnicas de contagem — Combinações, 79
12. Ilustrações adicionais, 85
Exercícios, 90

DISTRIBUIÇÃO DE PROBABILIDADES, 97

4

1. Introdução, 97
2. Variáveis aleatórias, 99
3. Propriedade das distribuições de probabilidade, 102
4. Valor esperado, 107
5. Variáveis contínuas, 112
6. Ilustrações adicionais, 115
Exercícios, 119

ALGUMAS DISTRIBUIÇÕES ESPECIAIS DE PROBABILIDADE, 121

5

1. Distribuição binomial, 121
2. Propriedades da distribuição binomial, 130
3. Distribuição normal, 132
4. Aproximação normal à binomial, 139
5. Obtenção da distribuição binomial, 147
6. Ilustrações adicionais, 148
Exercícios, 150

AMOSTRAGEM, 156

6

1. Introdução, 156
2. Amostras aleatórias, 157
3. Estimadores não tendenciosos, 159
4. Distribuição de $\bar{x}$ como amostra de uma população normal, 161
5. Distribuição de $\bar{x}$ como amostra de uma população não normal, 167
6. Ilustrações adicionais, 170
Exercícios, 171

ESTIMATIVA, 175

7

1. Estimativas por ponto e por intervalo, 175
2. Estimativa de μ, 176
3. Aproximações, 182
4. Estimativas de p, 184
5. A distribuição t de Student, 188
6. Ilustrações adicionais, 191
Exercícios, 195

TESTES DE HIPÓTESE, 201

8

1. Dois tipos de erro, 201
2. Teste de média, 207
3. Teste de proporção, 214
4. Teste da diferença de duas médias, 217
5. Teste de diferença de duas proporções, 222
6. Método de pequenas amostras, 224
7. Ilustrações adicionais, 228
Exercícios, 233

CORRELAÇÃO E REGRESSÃO, 241

9

1. Correlação linear, 241
2. Interpretação de r, 247
3. Fidedignidade de r, 248
4. Regressão linear, 250
5. Mínimos quadrados, 252
6. A falácia de regressão, 257
7. Erro-padrão da estimativa, 259
8. Método das pequenas amostras, 262
9. Ilustrações adicionais, 264
Exercícios, 268

TÓPICOS ADICIONAIS

A DISTRIBUIÇÃO QUI-QUADRADO, 275

10

1. Introdução, 275
2. A distribuição qui-quadrado, 277
3. Limitações do teste de qui-quadrado, 281
4. Tabelas de contingência, 282
Exercícios, 287

ANÁLISE DE VARIÂNCIA, 292

11

1. Introdução, 292
2. Uma variável de classificação, 293
3. A distribuição de F, 296
4. Notação ANOVA, 297
5. Duas variáveis de classificação, 301
6. Fórmulas de cálculo, 307
Exercícios, 310

TESTES NÃO-PARAMÉTRICOS, 313

12

1. Introdução, 313
2. Teste da mediana, 314
3. Teste da diferença de duas medianas, 316
4. Coeficiente de correlação de postos, 319
Exercícios, 320

PREFÁCIO À QUARTA EDIÇÃO

Este livro foi elaborado para cursos de um ou dois semestres e destina-se ao estudante cuja base matemática se restringe à álgebra introdutória. Muitos departamentos de métodos quantitativos oferecem cursos deste tipo; em algumas escolas, os próprios departamentos interessados em aplicações estatísticas se prontificam a dar o curso.

A edição atual difere da anterior principalmente no tocante a ilustrações e problemas que foram ampliados com numerosas e interessantes referências a questões sociais atuais. Os dados da maioria destes problemas foram obtidos de revistas de medicina ou de outras referentes a ciências sociais e ciências da natureza. A única mudança notável na organização do livro foi a introdução de um novo capítulo sobre regressão múltipla com a conseqüente eliminação desse assunto do capítulo sobre correlação e regressão.

A última seção de cada capítulo contém exemplos adicionais dos métodos apresentados. Pode ser usada pelo estudante como base de revisão de todo o capítulo ou como fonte de outros exemplos para as seções individuais do capítulo. Para facilitar este último tipo de uso, as ilustrações vêm com o próprio número da seção. A última seção de cada conjunto de problemas foi intitulada *Geral* porque as soluções destes problemas podem envolver várias das técnicas do capítulo. São problemas muito úteis como exercícios de revisão.

O material dos nove primeiros capítulos preenche as necessidades normais de um curso de um semestre. Os capítulos restantes, chamados *Tópicos Adicionais* têm material para se ampliar o curso básico, se houver tempo, ou para se organizar um curso de dois semestres. Estes cinco capítulos são independentes entre si e o

instrutor pode escolher aquele que lhe chamar a atenção. Como muitos capítulos foram aumentados, especialmente os cinco últimos; como há grande número de problemas novos, existe material suficiente para um curso de dois semestres. Muitos destes novos problemas são de interesse atual, o que justifica que se gaste muito tempo discutindo-os, antes de se avançar na matéria.

Muitos instrutores que adotaram as edições anteriores tiveram a gentileza de me enviar sugestões para a melhoria do livro. Adotei muitas destas sugestões. Mesmo não podendo incluir todas as idéias, apreciei muito a boa vontade dos que me tentaram ajudar a escrever um livro melhor.

Los Angeles, California
Junho 1975

Paul G. Hoel

Nota dos editores brasileiros

No final da obra encontram-se as soluções dos exercícios propostos em cada capítulo. Na edição original, estas soluções encontravam-se em manual separado. Na tradução para o português, após consultas dirigidas a vários especialistas da área, optamos pela incorporação das soluções ao texto.

1 A NATUREZA DOS MÉTODOS ESTATÍSTICOS

1 — INTRODUÇÃO

De um ponto de vista muito geral, métodos estatísticos são métodos de se tratar dados numéricos. Assim, métodos de coletar e analisar dados de qualquer empresa particular ou governamental podem ser considerados como pertencentes à estatística. Mas, tal definição é muito ampla para ser útil. É necessário restringir tanto a natureza dos dados quanto as razões de estudá-los, antes que tais métodos, com direito, sejam chamados estatísticos.

Os métodos estatísticos referem-se a dados obtidos de observações, na forma de medidas ou contagem, a partir da fonte de tais observações. Por exemplo, ao estudarmos o custo dos serviços médicos de uma cidade, escolheríamos uma pequena percentagem de habitantes da cidade e os interrogaríamos sobre seus gastos médicos. Ou, ao estudarmos a opinião pública sobre um problema polêmico debatido no Congresso, deveríamos escolher alguns eleitores dispersos pelo país e sondar suas opiniões sobre o problema.

Os estatísticos fazem tais observações com o intuito de tirar conclusões sobre o campo observado. Assim, os gastos médicos dos indivíduos escolhidos seriam usados para determinar o nível geral dos custos médicos na cidade. Igualmente, a finalidade da sondagem da pequena amostra de eleitores sobre o problema polêmico tem como fim determinar, com aproximação satisfatória, a opinião de todos os eleitores sobre o problema.

O conjunto de observações colhidas da fonte, com vistas à obtenção de informações sobre ela, chama-se *amostra,* e a fonte de tais observações chama-se *população.* Levando-se em conta o que dis-

cutimos antes, pode-se descrever o método estatístico como aquele método que tira conclusões sobre a população partindo da amostra. A palavra *estatística* é freqüentemente empregada como sinônimo de *método estatístico*. Um estudante que faz um curso de estatística está fazendo um curso sobre métodos estatísticos.

À primeira vista, tal definição pode parecer técnica e contrária à noção popular de estatística. Por exemplo, muitos homens de negócio consideram os métodos estatísticos como métodos de coletar e resumir fatos relacionados a negócios. O Governo Federal emprega muitos profissionais de estatística cuja principal obrigação é apresentar maneiras eficientes de coletar e sumariar diversos tipos de informações. Conforme a definição anterior de método estatístico, tais funcionários parecem não estar usando métodos estatísticos porque não aplicam a informação coletada no sentido de tirar conclusões sobre as fontes de informação. Tal modo de pensar, contudo, não leva em consideração o fato de que tais informações são colhidas para que outros trabalhem sobre elas, na busca de tais conclusões. Homens de negócio não coletam dados econômicos apenas para administrar a informação obtida. Esperam usá-los para tomar decisões, e chegando ou não chegando a conclusões relativas a fontes de informação, permanece o fato de que realmente decidem baseados em amostras.

A parte da estatística referente à coleta e tabulação de dados é conhecida, geralmente, como *estatística descritiva*. A parte referente a conclusões sobre as fontes de dados chama-se *inferência estatística*. Como o objetivo final é fazer inferências, isto é, tirar conclusões, a parte descritiva deveria ser considerada como uma espécie de introdução ao corpo principal. O uso dos métodos estatísticos, especialmente nas ciências sociais e biológicas, aumentou consideravelmente nas últimas décadas. Tais métodos mostraram-se também muito úteis nos vários ramos das ciências físicas e da engenharia. Devido a este interesse variado e amplo, tais métodos desenvolveram-se rapidamente e cresceram em diversidade e complexidade. Contudo, muitas das técnicas mais importantes são muito simples e são as mesmas para os vários ramos de aplicação. Alguns destes métodos universais serão estudados neste livro.

Serão úteis tanto a estudantes que desejam entender como são manipulados dados experimentais simples, quanto àqueles que necessitam este tipo de base para trabalhos mais avançados.

2 — ILUSTRAÇÕES

Esta seção descreve alguns problemas do tipo daqueles que os métodos estatísticos se destinam a resolver. Não cobre a ampla classe de problemas capazes de serem resolvidos por métodos estatísticos, mas ilustra alguns dos problemas mais simples que podem ser resolvidos pelo recurso a métodos desenvolvidos neste livro. Um problema é de interesse acadêmico, os demais são problemas típicos da vida real.

a) O prefeito de uma grande cidade deseja saber suas chances de ganhar as eleições primárias de seu partido para uma vaga no Congresso. Para conseguir tal informação, uma organização destinada a promover campanhas políticas concorda em obter uma amostragem de eleitores estaduais para determinar a popularidade do prefeito. Usando métodos estatísticos, tal organização pode saber qual o tamanho da amostra necessária para se estimar, com o necessário grau de precisão, a percentagem de eleitores favoráveis ao prefeito.

b) Uma equipe de pesquisa médica desenvolveu um novo tipo de soro que ela espera impedirá determinada doença infantil. Deseja testar o soro. Para ajudar os pesquisadores a levar a cabo a experiência, certo sistema escolar de uma grande cidade concordou em inocular este soro em metade das crianças de algumas classes. Conserva-se durante o ano seguinte o registro de todas as crianças destas classes. Com base na percentagem das crianças que contraíram a doença durante o ano, tanto do grupo inoculado como do restante, é possível determinar por métodos estatísticos se o soro é realmente benéfico ou não.

c) Certa empresa industrial está preocupada com o grande número de acidentes que ocorre em sua fábrica. Tentando descobrir as várias causas de tais acidentes, um pesquisador considera fatores relacionados às várias horas do dia. Coleta informações sobre o número de acidentes que ocorrem durante as várias horas de trabalho, e recorrendo a métodos estatísticos será capaz de mostrar que a taxa de acidentes aumenta, realmente, de manhã e à tarde. Estudos estatísticos posteriores revelarão os fatores mais responsáveis pelos acidentes.

Alguém poderia dizer que não é preciso recorrer a métodos estatísticos para se resolver um problema como este. Tudo que se necessita fazer é calcular as porcentagens e estudá-las para se saber o que está acontecendo. Se tivermos grande número de dados adequadamente selecionados, tais decisões estarão, freqüentemente, corretas. Mas os altos custos das coletas de dados força-nos a trabalhar com pequenas amostras, e é, precisamente, em tais situações, que necessitamos dos métodos estatísticos para tirarmos conclusões corretas.

d) Uma empresa de marketing deseja determinar o tamanho de um orçamento de propaganda que maximizará o lucro de um de seus produtos. Decide trabalhar em 3 áreas de mercado

bem separadas e levar a cabo programas promocionais variando propaganda e custos que julga adequados para o produto.

Ao fazer isso, necessita de métodos estatísticos sobre otimização para planejar o estudo e estimar o melhor tamanho do orçamento a usar.

e) Um instrutor de estatísticas elementar está tendo dificuldade em convencer alguns de seus estudantes de que as chances de se ganhar numa roleta são as mesmas para quem acabou de ganhar algum dinheiro e para quem perdeu seguidamente. Com o intuito de convencê-los disso ele e vários estudantes realizaram o seguinte experimento numa roleta de Las Vegas: A máquina é acionada por uma hora ou até que acabem os recursos do instrutor ou dos alunos, independentemente da ordem. Registra-se o número de ganhos e perdas que ocorrem, imediatamente após cada lance ganho, juntamente com os montantes ganhos; e também o número de ganhos e perdas e os montantes ganhos, imediatamente após uma série de, digamos, cinco perdas. Dispondo de dados deste tipo, o instrutor poderia aplicar métodos estatísticos para convencer os céticos de seu acerto na matéria. Como uma série de lances de má sorte dificulta a demonstração deste acerto, a não ser que se jogue durante muito tempo, o instrutor deveria vir com muito dinheiro. Um experimento deste tipo convenceria também os estudantes de que as roletas se destinam a tirar dinheiro de pessoas ingênuas.

3 — ESTIMATIVAS E TESTES DE HIPÓTESE

Uma análise das ilustrações precedentes mostra que elas pertencem ao campo da estatística porque todas buscam tirar conclusões sobre populações de objetos ou de indivíduos e se propõem conseguir isso com base em amostragens populacionais.

Pode-se também observar que tais problemas caem em duas categorias gerais. Referem-se ou à estimativa de alguma propriedade da população ou a testes de hipótese sobre a população. A primeira ilustração, por exemplo, refere-se à estimativa da percentagem de eleitores do estado que favorecem o prefeito. A segunda ilustração trata de testar a hipótese segundo a qual a porcentagem de crianças inoculadas que contraiu a doença é a mesma que a porcentagem de crianças não inoculadas. A terceira ilustração considera o problema de testar a hipótese segundo a qual a taxa de acidentes de uma população é constante, ao longo do dia. A quarta ilustração refere-se à estimação de lucro como função do montante gasto na propaganda do produto. A quinta ilustração testa a hipótese de que o montante médio de dinheiro ganho numa roleta, após uma série de lances perdidos, é o mesmo após um lance ganho.

A maioria dos métodos estatísticos que serão explicados neste livro trata destes dois tipos de problemas, a saber: estimativa de

propriedades ou testes de hipóteses sobre populações. Embora haja outros tipos de conclusões ou decisões que podem referir-se a populações na base de amostras, a grande maioria dos problemas tratados por estatísticas cai dentro destas duas categorias e, portanto, compõe a maior parte do material deste livro.

4 — PROBABILIDADE

No problema de estimar a porcentagem de eleitores que favorecem um candidato ou um problema, a solução consistirá numa porcentagem baseada sobre a amostra e na proposição de exatidão da estimativa, geralmente em termos como, por exemplo, "há 95% de probabilidade de que a estimativa esteja errada em apenas 3%". Do mesmo modo, em problemas que envolvem testes de hipóteses, a decisão de aceitarmos ou rejeitarmos a hipótese estará baseada em certas probabilidades.

É necessário usarmos probabilidades em tais conclusões porque uma conclusão baseada numa amostra apresenta informação incompleta sobre a população e, portanto, não pode ser feita com certeza. A maior ou menor probabilidade ligada a uma conclusão representa o grau de confiança que se deve ter na verdade da conclusão. A idéia e as regras básicas de probabilidade serão estudadas em capítulo posterior. Por enquanto tratá-la-emos de um ponto de vista intuitivo. Assim, a proposição "há 95% de probabilidade de que a estimativa esteja errada em menos de 3%" significa que 95% da proposição feita pelo estatístico é válida e 5% não o é. Ao estudarmos métodos estatísticos, descobriremos logo que a probabilidade é o instrumento básico destes métodos.

A probabilidade é um assunto extremamente interessante, mesmo para aqueles que têm pouco gosto por matemática ou métodos quantitativos. Muitas pessoas gostam de eventos ligados a probabilidade, mesmo não a estudando. De outra maneira, como poderíamos explicar o grande número de pessoas que jogam em corridas de cavalos, loterias etc...? Pode mesmo ser sua falta de sentido de probabilidade que os anima a jogar. Seja como for, a probabilidade é usada consciente ou inconscientemente por qualquer que decida em situações de incerteza e qualquer estudante que deseje ser bem educado ou se comportar racionalmente deveria ter algum conhecimnto de probabilidade.

5 — ORGANIZAÇÃO

O estudo dos métodos estatísticos discutidos nas seções precedentes considerará primeiro as propriedades das amostras e depois as propriedades das populações. Como foi indicado na seção 1, tais estudos constituem a parte descritiva da estatística. Depois, será possível considerar os dois problemas básicos da inferência estatística, isto é, problemas de estimativa e testes de hipótese. Isto significa que, para qualquer tipo de problema, os dados amostrais sempre serão estudados primeiro, antes que se faça qualquer tentativa para se introduzir uma população teórica da qual poder-se-ia originar a amostra. No Capítulo 2, começaremos a estudar as propriedades dos dados amostrais e introduziremos depois algumas populações teóricas básicas. Neste estágio teórico, a probabilidade aparecerá.

2

DESCRIÇÃO DE DADOS AMOSTRAIS

1 — INTRODUÇÃO

Como o propósito deste capítulo é estudar propriedades de amostras tiradas de populações, seria necessário primeiro concordarmos sobre como colher tais amostras, porque as propriedades desejadas podem muito bem depender do método empregado. Suponhamos, por exemplo, que um repórter de um jornal estudantil foi designado para determinar a porcentagem de estudantes com empregos de meio-período. Ele poderia tentar obter esta porcentagem arrolando os 100 primeiros estudantes que encontrasse diante do diretório. Tal método de amostragem, contudo, provavelmente não ofereceria uma estimativa válida da porcentagem da população porque, geralmente, os estudantes que se preocupam muito com diretórios são "flanadores do campus", "borboletas sociais". Raramente são pessoas que trabalham. Sem dúvida, o repórter faria muito melhor se escolhesse cegamente 100 cartões de matrícula e os tomasse como a amostra selecionada.

É muito complexo o problema de como selecionar a amostra de uma população, de modo a se poder tirar conclusões válidas sobre ela.

Este problema é discutido de modo extenso no Capítulo 6. Nesta discussão advoga-se e se justifica um método de amostragem chamado amostragem aleatória. Antecipando de certo modo este assunto material, damos aqui uma introdução a este tipo de amostragem. Na sua forma mais simples, quando se vai selecionar um único indivíduo de uma população, diz-se que a amostragem é aleatória se cada um dos membros da população tem a mesma chance de ser escolhido.

Freqüentemente, para se obter tais amostras empregam-se técnicas usadas em jogos de azar. Por exemplo, numa festa social em que se vai sortear um grande prêmio, colocam-se geralmente todos os cartões numerados num recipiente qualquer escolhendo-se então um indivíduo que com os olhos vendados vá tirar os cartões após tê-los misturado bem.

Se, digamos, três indivíduos devem ser escolhidos de uma população, a amostra será aleatória se cada grupo possível de três indivíduos da população tiver a mesma chance de ser escolhido. O mesmo artifício usado na escolha de um indivíduo poderia ser usado na escolha de três indivíduos. Experiências com tais artifícios mostraram que cada indivíduo, ou grupo de indivíduos, é selecionado aproximadamente o mesmo número de vezes, quando o experimento é suficientemente grande, e que, portanto, eles se conformam à exigência de não se mostrar favoritismo. Se selecionássemos uma amostra de estudantes de um corpo estudantil, escolhendo sempre o 10.º cartão das filas de inscrição, estaríamos, sem dúvida, aplicando um método prático e satisfatório de seleção aleatória, se a informação desejada sobre o indivíduo não estiver relacionada à sua posição alfabética na fila de cartões.

Entre as várias propriedades desejadas, as amostras aleatórias tendem a refletir em miniatura as populações das quais são tiradas. Isto faz com que, por exemplo, se o experimento de selecionar um estudante aleatoriamente — de uma população estudantil — for repetido várias vezes, ao calcularmos a percentagem de estudantes que trabalham em tempo parcial, tal percentagem se aproximará cada vez mais da percentagem real da população, com a continuação dos experimentos. Outros tipos de amostragem onde entra o julgamento pessoal geralmente não possuem tal propriedade.

Como ilustração desta propriedade de amostras aleatórias representando a população amostrada, levou-se a cabo um experimento para uma população construída artificialmente. Suponhamos que um grande número de indivíduos pudesse ser classificado em três grupos. Por exemplo, poderia ser na base da idade: pessoas de menos de 30 anos, pessoas entre 30 e 50 inclusive e pessoas com mais de 50 anos. Suponhamos, além disso, que as proporções dos indivíduos nestes três grupos sejam de 3/6, 2/6 e 1/6 respectivamente. A amostra aleatória de tal população poderia ser simulada fazendo-se amostras de uma população de 6 cartas de baralho, sendo, por exemplo: 3 ases, 2 dois e 1 três, fazendo-se a amostragem de maneira adequada. Após cada retirada de carta, registra-se o número e repõe-se a carta retirada. Embaralham-se novamente as cartas e faz-se nova retirada. O ato de embaralharmos e devolvermos a carta retirada assegura-nos que a proporção populacional permanece a mesma e não ocorrerá

favoritismo. Se a população original for muito grande, a remoção de alguns indivíduos não terá efeito apreciável sobre as proporções da população. O experimento anterior foi efetuado 700 vezes em conjunto de 100. Os resultados acumulados foram reduzidos a porcentagens, arredondados depois que se obtinha cada amostra adicional de 100. Estas porcentagens amostrais aleatórias, mostradas no Quadro 1, devem ser comparadas com as porcentagens populacionais de 50,33 1/3 e 16 2/3 respectivamente, para vermos o grau de fidelidade na representação amostral de uma população.

QUADRO 1

Tamanho da amostra	100	200	300	400	500	600	700
Ases	44	47	45,7	47,8	49,4	49,6	50,0
Dois	35	34	35,0	33,2	32,8	32,8	32,9
Três	21	19	19,3	19,0	17,8	17,5	17,1

A maneira pela qual estas percentagens amostrais convergem para as percentagens populacionais correspondentes são mais facilmente visualizadas de maneira geométrica, como mostra a Figura 1. Parecem convergir melhor para as percentagens populacionais à medida que aumenta o tamanho da amostra.

Daqui por diante, sempre que quisermos obter uma amostra, suporemos que será obtida por um método amostral aleatório, mesmo quando a palavra aleatória não for usada explicitamente.

Voltemos agora ao problema do estudo das propriedades das amostras populacionais. Neste sentido, embora a palavra população pareça referir-se a grupos de indivíduos humanos, em estatística ela se refere a um grupo de indivíduos humanos ou não, ou a objetos de qualquer espécie. Ao estudarmos amostras ou populações supõe-se, freqüentemente, que o interesse esteja centrado numa propriedade particular única dos membros da população. Assim, alguém poderia estar interessado na propriedade "peso" de um grupo de universitários ou na cor dos olhos de uma população de insetos ou na porcentagem de ferro de uma população de meteoritos. Tira-se de uma população amostras de indivíduos, mas mede-se ou conta-se depois a característica procurada. A propriedade medida ou contada será, geralmente, expressa pela letra X. Assim, X poderia representar o peso do estudante ou o número de crianças numa família ou os gastos médicos anuais de um receituário médico.

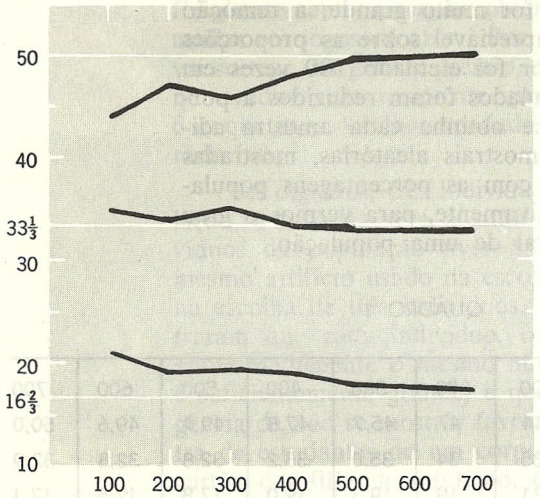


FIGURA 1. Resultados amostrais aleatórios.

Como ilustração do modo de se proceder no estudo de amostras, consideremos como agiria o departamento de educação física de uma universidade, se quisesse determinar que os estudantes internos do sexo masculino são universitários típicos com respeito a características físicas. Seria preciso, sem dúvida, comparar em tal estudo, para termos um referencial, a distribuição de peso dos estudantes internos e externos. Pesando cada estudante do campus, teríamos sem dúvida a distribuição desejada. Mas isto seria muito trabalhoso, numa grande escola onde tal informação não fosse exigida no registro de matrícula. A informação desejada poderia ser obtida com suficiente exatidão e de modo muito mais facil se estudarmos as distribuições de peso de amostras de estudantes internos e externos.

Suponhamos então que se obtenha uma amostra aleatória de 120 estudantes da população interna. Como a preocupação, aqui, é saber o que fazer com as amostras, a população de alunos externos pode ser ignorada, no caso (seria tratada da mesma maneira que a população interna). Vamos supor, também, que os pesos destes 120 estudantes tenham sido registrados e arredondados e que eles variem de 110 libras * para os mais leves a 218 para os mais pesados.

* Uma libra equivale a 453,592 g. (N. do R.)

É muito difícil olhar 120 medidas e obter uma idéia razoavelmente exata da distribuição de tais medidas. Com o propósito de obtermos uma melhor idéia da distribuição de peso dos 120 estudantes, será conveniente condensarmos de algum modo os dados, classificando as medidas em grupos. Poderemos então traçar o gráfico da distribuição modificada e aprendermos mais sobre o conjunto original das 120 medidas. Tal condensação será útil, também, para simplificarmos os cálculos de várias médias que devem ser avaliadas, especialmente se não houver possibilidade de se recorrer a métodos eletrônicos de computação. Assim, a finalidade da classificação dos dados é facilitar a extração de certos tipos de informações úteis, a partir dos dados.

As medidas de peso consideradas aqui são um exemplo de observações feitas sobre o que se chama *variável contínua*. Aplica-se esta denominação a variáveis, como comprimento, peso, temperatura e tempo que podem assumir qualquer valor numa determinada escala de valores. Assim o peso de um estudante dentro da escala de 140 — 150 libras pode assumir qualquer valor dentro desta escala. Variáveis tais como o número de acidentes automobilísticos por dia, número de insetos que morrem sob a ação de inseticidas ou número de filhos de uma família são exemplos que costumamos chamar *variáveis discretas*. Para os propósitos deste livro, as variáveis discretas podem ser consideradas como variáveis cujos valores possíveis são inteiros. O que está em jogo, no caso, é a contagem e não a medição.

Como os artifícios de medida são de exatidão limitada, as medidas da vida real são realmente discretas, na natureza, e não contínuas. Mas isto não impede que se pense em tais variáveis como sendo contínuas. Embora os pesos tenham sido registrados e arredondados, devem ser considerados como valores de uma variável contínua. Tais valores foram arredondados para o inteiro mais próximo. Quando se registra um peso de, digamos, 152 libras, presume-se que o peso real estará entre 151,5 e 152,5 libras.

2 — CLASSIFICAÇÃO DE DADOS

O problema de classificação de dados amostrais só surge para variáveis contínuas. Como exemplo, consideremos o problema da medida de peso dos 120 estudantes internos. O que se precisa fazer é escolher um conjunto de intervalos de classe e depois colocar cada peso em sua própria classe. Por exemplo, se escolhermos um intervalo de classe de 10 libras, todos os pesos entre 130 e 139 inclusive deveriam ser colocados na mesma classe. Como tais pesos foram

arredondados, este intervalo de classe especial estender-se-ia de 129,5 a 139,5. Os valores 129,5 e 139,5 chamam-se *limites de classe* desta classe. Após determinar as fronteiras de classe é fácil colocar cada medição em seu próprio intervalo de classe. Depois de completarmos a listagem e registrarmos o número de pesos de cada intervalo de classe, diz-se que os dados foram classificados numa *tabela de freqüência.*

QUADRO 2

Limites de Classe	Freqüências
109,5-119,5	1
119,5-129,5	4
129,5-139,5	17
139,5-149,5	28
149,5-159,5	25
159,5-169,5	18
169,5-179,5	13
179,5-189,5	6
189,5-199,5	5
199,5-209,5	2
209,5-219,5	1

O Quadro 2 ilustra o resultado de tal processo de classificação para um conjunto de 120 pesos do tipo em consideração.

É conveniente usarmos de 10 a 20 intervalos de classe. A razão de tal escolha tornar-se-á clara na próxima seção.

3 — REPRESENTAÇÃO GRÁFICA

As distribuições de freqüência são mais fáceis de visualizar se as representarmos graficamente. Para variáveis do tipo contínuo usa-se geralmente o gráfico chamado *histograma.* Pode também ser usado para variáveis discretas. O histograma da distribuição de freqüência do Quadro 2 aparece na Figura 2. Os limites de classe do Quadro 2 são marcados no eixo dos x, começando e terminando nos pontos convenientes. A freqüência correspondente a cada intervalo de classe é representada pela altura do retângulo cuja base é o intervalo, em

questão. O eixo vertical é, portanto, o eixo de freqüência. Os histogramas são gráficos particularmente úteis para trabalhos posteriores, quando se introduzem as distribuições de freqüência de populações.

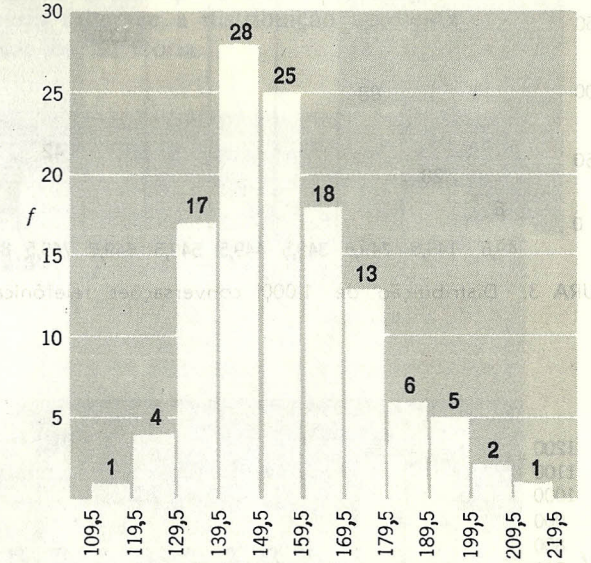


FIGURA 2. Distribuição dos pesos de 120 estudantes

O histograma da Figura 2 é típico de muitas distribuições de freqüência obtidas de dados encontrados na natureza e na indústria. Tais distribuições variam geralmente de uma distribuição grosseiramente em forma de sino, como a da Figura 3, a uma distribuição que apresenta a metade direita em forma de sino, como a da Figura 5.

Nestes dois gráficos registra-se o valor médio de cada intervalo sobre o eixo dos x, no lugar dos limites de classe. Isto é conveniente para o uso futuro desses histogramas. Uma distribuição como a da Figura 5 diz-se assimétrica. A assimetria é medida em relação ao eixo vertical.

Se o histograma tem uma cauda direita longa e uma cauda esquerda curta, diz-se que é assimétrico à direita. Na situação inversa, será assimétrico à esquerda. Quanto maior o desequilíbrio tanto

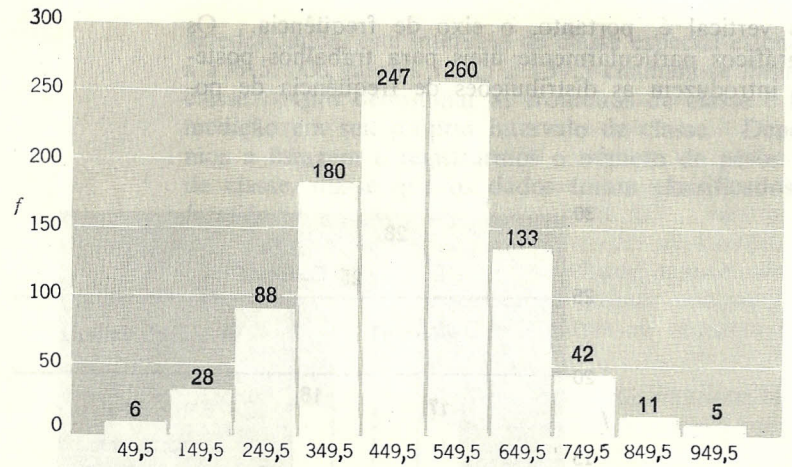


FIGURA 3. Distribuição de 1 000 conversações telefônicas em segundos.

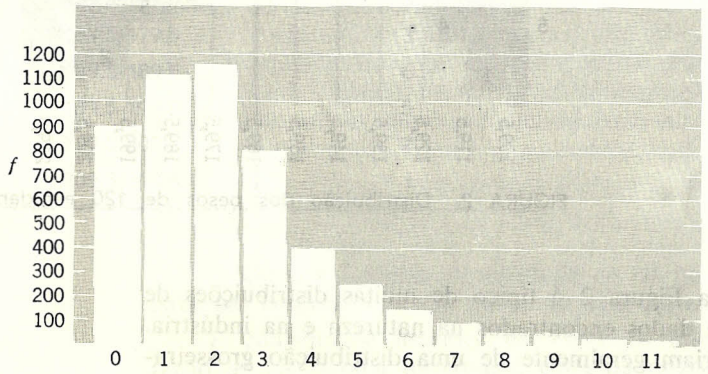


FIGURA 4. Distribuição do número de crianças nascidas de 5 000 mulheres casadas há 20 anos ou mais.

maior o grau de assimetria. Por exemplo, sabemos que as seguintes variáveis têm distribuições que possuem formas do tipo como acabamos de ver, com graus de assimetria crescentes, estatura, muitas medidas industriais, várias medidas biológicas lineares, peso, idade de casamento, idade mortal para certas doenças, e saúde. As figuras 3, 2, 4, e 5 mostram os histogramas para quatro distribuições típicas com graus crescentes de assimetria.

4 — DESCRIÇÃO ARITMÉTICA

A principal razão para classificar os dados e desenhar o histograma da tabela de freqüências resultante é determinar a natureza da distribuição. Muito da teoria explanada em capítulos posteriores requer que a distribuição se assemelhe ao tipo observado na Figura 3; conseqüentemente é necessário saber se a distribuição apresenta este tipo, antes de tentar aplicar-lhe tal teoria.

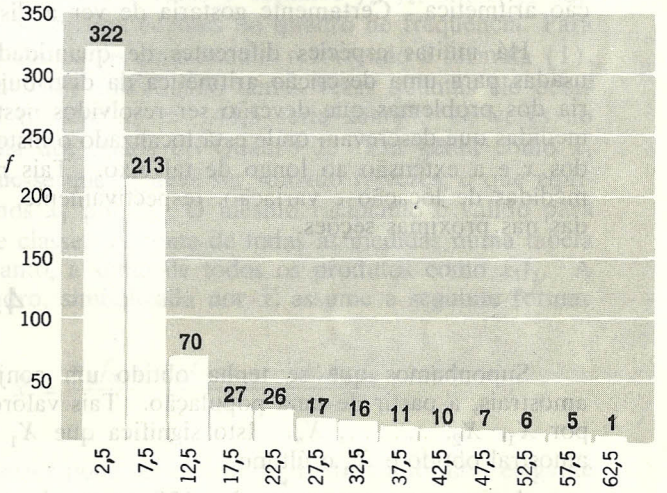


FIGURA 5. Distribuição de 727 mortes provocadas por febre escarlate classificadas conforme a idade.

Embora o histograma forneça grande número de informações gerais relativas à distribuição de um conjunto de medidas amostrais, deve-se obter informações mais úteis e precisas para o estudo da distribuição a partir de uma descrição aritmética da distribuição. Por exemplo, se o histograma de pesos para uma amostra de estudantes internos for disponível para comparação com o histograma da amostra do internato, será difícil constatar, exceto em termos muito gerais, como diferem as distribuições. Ao invés de comparar as duas distribuições de peso na totalidade, é suficiente comparar os pesos médios e a variação, em peso, dos dois grupos. Tais quantidades descritivas são ditas aritméticas porque conduzem a números, em contraste com o histograma que é de natureza geométrica.

A natureza de um problema estatístico determina largamente se algumas propriedades aritméticas simples da distribuição seriam sufi-

cientes para descrevê-las satisfatoriamente. A maioria dos problemas encontrados neste livro são do tipo que requer apenas algumas propriedades da distribuição para solucioná-lo. Para problemas mais avançados tal condensação de informações fornecida pela distribuição pode não ser suficiente. A situação aqui é semelhante à que surge quando se discutem problemas relacionados com roupas de mulher e concurso de beleza. Um vendedor de vestidos por exemplo, pode ficar satisfeito se lhe disserem que uma moça tem 5 pés * e 7 polegadas ** de altura e tem medidas de 36-24-36 polegadas. Mas um juiz de concurso de beleza dificilmente se satisfaria com tal descrição aritmética. Certamente gostaria de ver a distribuição inteira.

Há muitas espécies diferentes de quantidades que podem ser usadas para uma descrição aritmética da distribuição. Para a maioria dos problemas que deverão ser resolvidos neste livro, bastará ter medidas que descrevam onde está localizado o histograma sobre o eixo dos x e a extensão ao longo de tal eixo. Tais medidas chamam-se medidas de locação e variação, respectivamente. Serão elas discutidas nas próximas seções.

4.1 — Média

Suponhamos que se tenha obtido um conjunto de n valores amostrais, a partir de uma população. Tais valores são simbolizados por $X_1, X_2, \ldots\ldots, X_n$. Isto significa que X_1 é o primeiro valor amostral obtido e X_n o último.

Assim, para a amostra dos 120 pesos, X_1 seria o peso do primeiro aluno escolhido e X_{120} o peso do último aluno. A média deste conjunto de números, simbolizada por $\bar{X}$ é dada pela fórmula:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{1}$$

Como há outras médias também usadas em estatística, é necessário darmos a $\bar{X}$ um nome especial. É a *média aritmética* ou por brevidade, a média. Consideremos agora o problema de calcularmos a média quando os dados foram classificados numa tabela de freqüência, como as medidas do Quadro 2. Para tais dados presume-se que cada um dos pesos originais foi substituído pelo peso do ponto médio do intervalo de classe no qual a medida cai. Este valor do ponto

° Um pé equivale a 30,480 cm. (N. do R.)
°° Uma polegada equivale a 25,4 mm. (N. do R.)

médio chama-se *ponto de classe* do intervalo. Assim, todos os estudantes cujos pesos estão no intervalo de classe de 139,5 a 149,5 deveriam receber o peso de 144,5 porque este é o ponto médio do intervalo. Podemos observar no Quadro 2 que há 28 estudantes cujos pesos poderiam ser determinados como 144,5 após classificação. Como resultado, para os dados classificados, os únicos valores de x que surgem no cálculo da média da amostra são os valores dos pontos de classe dos intervalos. Suponhamos que haja k intervalos de classe na tabela de freqüência. Façamos $x_1, x_2, \ldots\ldots, x_k$ denotar os valores do ponto médio desses k intervalos e façamos $f_1, f_2, \ldots, f_k$ significar as freqüências correspondentes no quadro de freqüência. Para se calcular a média dos dados classificados recorrendo à fórmula (1), bastará somar cada valor classificado tantas vezes quantas ele ocorrer. Por exemplo, o valor $x_4 = 144{,}5$ seria somado 28 vezes ao calcularmos a média dos dados do Quadro 2. Em geral, como x_1 ocorre f_1 vezes segue-se que x_1 deve ser somado f_1 vezes, o que equivale a multiplicarmos x_1 por f_1. O mesmo raciocínio é válido para os outros pontos de classe. A soma de todas as medidas numa tabela classificada é, portanto, a soma de todos os produtos como $x_1 f_1$. A média para tal quadro, simbolizada por $\bar{x}$, assume a seguinte forma:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n} \tag{2}$$

A fórmula anterior pode ser escrita de modo muito mais elegante se usarmos o símbolo Σ (sigma maiúsculo grego). Como este símbolo aparece, também, em outras fórmulas, seria conveniente travarmos conhecimento com ele, desde já. É um mero símbolo, que nos diz que devemos somar todos os valores da quantidade escritos após o símbolo.

Assim, a fórmula (1) é escrita assim:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3}$$

O símbolo Σ tem n sobre si e $i=1$ em baixo para indicar que todos os números como X_i devem ser somados, começando com $i=1$ e terminando com $i = n$. O símbolo X_i simboliza o valor amostral i-ésimo, onde i é qualquer inteiro de 1 a n. Em palavras este símbolo diz simplesmente "some todos os valores dos Xs, começando com o primeiro ($i=1$) valor amostral e terminando com o último valor

amostral ($i=n$)". A fórmula (2) pode ser escrita, de maneira condensada, assim:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} \tag{4}$$

Substituindo-se cada medida de uma amostra por seu ponto de classe, ter-se-á geralmente um valor da média que difere ligeiramente da média das medidas originais; portanto, o valor da média dado pela fórmula (4) é somente uma aproximação do valor original dado pela fórmula (3); todavia, a diferença é comumente tão pequena que pode ser ignorada na maior parte dos problemas estatísticos.

Como, para as duas próximas seções, teremos necessidade do valor de $\bar{x}$ dos dados do Quadro 2, efetuaremos seu cálculo aqui, recorrendo à fórmula (4).

QUADRO 3

x_i	f_i	$x_i f_i$
114,5	1	114,5
124,5	4	498,0
134,5	17	2.286,5
144,5	28	4.046,0
154,5	25	3.862,5
164,5	18	2.961,0
174,5	13	2.268,5
184,5	6	1.107,0
194,5	5	972,5
204,5	2	409,0
214,5	1	214,5
Totais	120	18.740,0

$$\bar{x} = \frac{18.740,0}{120} = 156,2$$

Agora pode-se mostrar que o valor numérico da média representa o ponto sobre o eixo dos x onde uma folha de metal, com a forma do histograma, balançaria como se estivesse sobre o fio de uma faca. Como este ponto de equilíbrio situa-se, geralmente, próximo ao meio da base do histograma, segue-se que $\bar{x}$ dá uma idéia bastante boa do local onde o histograma está centrado. Assim a média ajuda a descrever uma distribuição de freqüência, ao dizer onde se localiza ao longo do eixo dos x, o histograma de distribuição. Por isso, muitas vezes é chamado *medida de localização*.

Para a distribuição de freqüência do Quadro 2, cujo histograma aparece na Figura 2, os cálculos dão o valor de $\bar{x} = 156,2$ como vimos no Quadro 3. Um exame da Figura 2 nos mostrará certamente que o histograma deve oscilar sobre o fio da faca nas proximidades de 156. Assim o valor de $\bar{x}$ aqui dá uma boa idéia de onde se localiza o histograma de distribuição, no eixo dos x.

A principal razão da classificação de um conjunto de medidas é obter informações relativas à natureza da distribuição da variável X, em estudo. Portanto, mesmo que os dados tenham sido classificados, a média deveria ser calculada por meio da fórmula (3) e não pela fórmula (4), se o conjunto original de medidas for ainda disponível.

Várias outras medidas de localização serão discutidas na seção 5.

4.2 — Desvio-padrão

O objetivo desta seção é introduzir uma quantidade que meça a amplitude de variação em torno da média, de um conjunto de medidas. Neste sentido, consideremos os dois conjuntos de medidas:

4, 5, 6, 7, 8, e 2, 4, 6, 8, 10

e seus histogramas, que aparecem na Figura 6.

A média de cada conjunto é 6. Como a diferença comum entre os pares consecutivos de medidas do segundo conjunto é 2 vezes

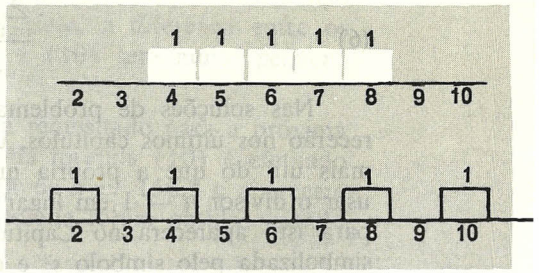


FIGURA 6. Duas distribuições artificiais.

a do primeiro conjunto (2 para 1) e tendo em vista também que a amplitude dos valores do segundo conjunto é o duplo da amplitude do primeiro conjunto (8 para 4), a maior parte das pessoas concordará em que o segundo conjunto de medidas varie duas vezes mais que o primeiro.

A medida mais simples da variação de um conjunto de medidas é, sem dúvida, a amplitude do conjunto, isto é, a diferença entre a medida maior e a menor do conjunto. Mas ela possui certas propriedades indesejáveis que serão discutidas na seção 5; por isso introduziremos outra medida mais útil.

Como a média foi escolhida como a medida desejada de localização de um conjunto de medidas, a medida de variação deveria medir a extensão dos desvios das medidas, com relação à média. Para dados não classificados correspondentes à fórmula (3), os desvios são $X_1 - \bar{X}$, $X_2 - \bar{X}$, ..., $X_n - \bar{X}$. Valores de X maiores que $\bar{X}$ produzirão desvios positivos e valores de X menores que $\bar{X}$ produzirão desvios negativos. Como se deseja apenas distâncias positivas da média, é preciso usar os valores absolutos de tais desvios ou usar seus quadrados. Como é difícil trabalhar com valores absolutos, geralmente eleva-se os desvios ao quadrado e tira-se a média deles.

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \tag{5}$$

Se se classificou um conjunto de medidas e se estará usando a fórmula (4) como medida de localização, então a média dos quadrados dos desvios é

$$\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n} \tag{6}$$

Nas soluções de problemas de inferência estatística, que aparecerão nos últimos capítulos, uma leve modificação desta medida é mais útil do que a própria medida. Esta modificação consiste em usar o divisor $n - 1$ em lugar do n destas fórmulas. A justificação para isto aparecerá no Capítulo 6. A quantidade resultante será simbolizada pelo símbolo s^2 e chama-se *variância amostral*. Assim, por definição, para dados não classificados

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1} \tag{7}$$

e para dados classificados

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n - 1} \tag{8}$$

Como a variância envolve os quadrados dos desvios, é um número em unidades elevadas ao quadrado. Em muitos problemas, é desejável que quantidades que descrevem uma distribuição possuam as mesmas unidades que o conjunto original de medidas. A média satisfaz este requisito, mas a variância não. Contudo, tirando-se a raiz quadrada positiva da variância, pode-se obter o efeito desejado. A quantidade chama-se desvio-padrão. Assim, por definição, o *desvio-padrão* para dados não classificados é

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}} \tag{9}$$

e para dados classificados é

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n - 1}} \tag{10}$$

Como no caso do cálculo da média, deveríamos usar a fórmula (9) ao invés da fórmula (10) no cálculo do desvio-padrão, se estiverem disponíveis os dados originais não classificados. Mas, a não ser no caso de uma classificação superexata, a diferença entre os valores de s obtidos das fórmulas (9) e (10) será muito pequena e poderá ser ignorada, na prática.

O valor de s para o Quadro 2 será requisitado para a próxima seção. Portanto, seu cálculo por meio da fórmula (10) é efetuado aqui. Os cálculos são baseados no valor de $\bar{x} = 156{,}2$ e aparecem no Quadro 4. Os quadros da quarta coluna foram obtidos usando o Quadro 1, do apêndice.

Muitas das calculadoras eletrônicas modernas (mesmo as manuais) fornecem o valor de s sem a necessidade de se calcular as

QUADRO 4

x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
114,5	1	−41,7	1.738,89	1.738,89
124,5	4	−31,7	1.004,89	4.019,56
134,5	17	−21,7	470,89	8.005,13
144,5	28	−11,7	136,89	3.832,92
154,5	25	− 1,7	2,89	72,25
164,5	18	8,3	68,89	1.240,02
174,5	13	18,3	334,89	4.353,57
184,5	6	28,3	800,89	4.805,34
194,5	5	38,3	1.466,89	7.334,45
204,5	2	48,3	2.332,89	4.665,78
214,5	1	58,3	3.398,89	3.398,89
Totais	120			43.466,80

$$s = \sqrt{\frac{43.466,80}{119}} = \sqrt{365,27} = 19,1$$

somas dos quadrados. Os cálculos do Quadro 4 foram feitos na suposição de que tais máquinas não estejam disponíveis.

4.3 — Interpretação do Desvio-padrão

Com o fito de observarmos como s mede a variação, consideremos mais uma vez os dois conjuntos de medidas correspondentes à Figura 6. Como $\bar{X} = 6$ para ambos os conjuntos, o cálculo de s para os dois conjuntos, através da fórmula (9) fornecerá os valores

$$s_1 = \sqrt{\frac{4 + 1 + 0 + 1 + 4}{4}} = \frac{\sqrt{10}}{2}$$

e

$$s_2 = \sqrt{\frac{16 + 4 + 0 + 4 + 16}{4}} = \sqrt{10},$$

O fato de o valor do desvio-padrão ser duas vezes maior no segundo conjunto que no primeiro é certamente uma propriedade satisfatória do desvio-padrão, se deve ser considerado como medida de variação.

A ilustração anterior, que indica que o desvio-padrão aumenta em tamanho com o aumento da variabilidade dos dados, não fornece nenhuma chave para o significado da grandeza do desvio-padrão.

Assim, se os valores do desvio-padrão nesta ilustração fossem $3\sqrt{10}$ e $6\sqrt{10}$, em vez de $\sqrt{10}/2$ e $\sqrt{10}$, a interpretação seria a mesma. Esta situação é semelhante à que ocorre quando dois estudantes comparam as notas em um teste. Um deles pode obter o dobro de pontos do outro, mas isto não revela quanto conhecimento absoluto da matéria cada estudante possui.

A fim de dar um significado quantitativo ao tamanho do desvio-padrão, é necessário antecipar certos resultados de trabalho posterior. Para um conjunto de dados obtidos por amostragem de um tipo particular de população, chamada população normal, mostraremos mais adiante que, quando a amostra é grande, o intervalo de $\bar{x} - s$ a $\bar{x} + s$ comumente inclui cerca de 68 por cento das observações e o intervalo de $\bar{x} - 2s$ a $\bar{x} + 2s$ usualmente inclui cerca de 95 por cento das observações. Uma amostra de população deste tipo comumente tem um histograma que lembra o histograma da Figura 2. Se a amostra aumenta de tamanho, o histograma tende a se aproximar da forma de um sino.

Como exemplo ilustrativo desta propriedade, consideremos os dados do Quadro 2. Os cálculos dos Quadros 3 e 4 forneceram os valores $\bar{x} = 156,2$ e $s = 19,1$. Os dois intervalos $(\bar{x} - s, \bar{x} + s)$ e $(\bar{x} - 2x, \bar{x} + 2s)$ tornaram-se portanto (137,1, 175,3) e (118,0, 194,4), respectivamente. Estes valores foram marcados com flechas verticais sobre o eixo x da Figura 7, que é o gráfico do histograma para a distribuição da freqüência do Quadro 2.

Calculando as porcentagens dos dados que estão dentro de cada um destes dois intervalos é necessário saber aproximadamente quantas observações em um dado intervalo ficam à direita e à esquerda de um ponto do mesmo. Para este fim fica convencionado que todas as observações de um intervalo se dispõem uniformemente ao longo deste intervalo.

Assim, para o segundo intervalo, presume-se que uma das quatro observações está no primeiro quarto do intervalo, a outra no segundo quarto do intervalo etc. Com este entendimento, o número de observações entre 137,1 e 175,3 é aproximadamente igual a $4+28+25+18+8 = 83$. Os valores de 4 e 8 são obtidos por interpolação de acordo com a convenção precedente sobre o modo de se disporem as observações. Por exemplo, 4 é obtido olhando a Figura 7 e percebendo que a distância de 129,5 e 137,1 é de 7,6 unidades, ao passo que a distância de 129,5 e 139,5 é de 10 unidades. Portanto admite-se que 76 por cento das 17 observações neste intervalo ficarão à esquerda de 137,1 e 24 por cento à direita. Desde que 24 por cento de 17 é igual a 4 arredondando para inteiro, segue-se que 4 das 17 observações neste intervalo ficarão à esquerda de 137,1. O

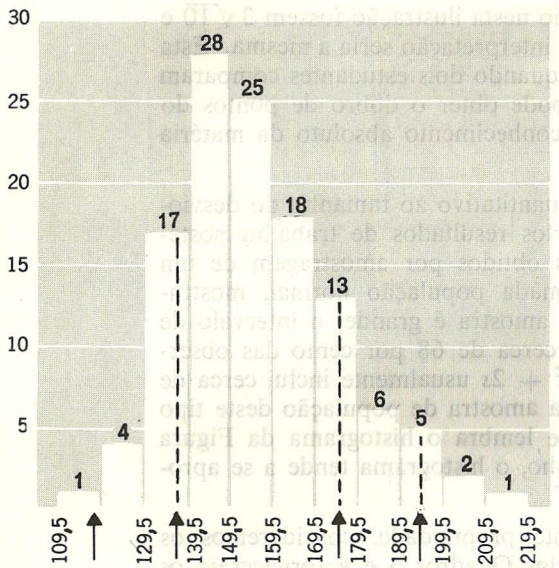


FIGURA 7. Histograma de distribuição de 120 pesos.

valor de 8 na soma anterior é obtido de maneira semelhante. Assim, a distância de 169,5 a 175,3 é de 5,8 unidades, enquanto a distância de 169,5 a 179,5 é de 10 unidades. Daí 58 por cento das 13 medidas deste intervalo deveriam ser tratadas como estando à esquerda do ponto 175,3. Como 58 por cento de 13 é 7,54, ou 8 arredondando, segue-se que 8 destas 13 observações deveriam estar à equerda do ponto 175,3. Como o número total de observações é 120 e como $83/120 = 0{,}69$, segue-se que aproximadamente 69 por cento das observações caem dentro do intervalo $(\bar{x} - s, \bar{x} + s)$. Cálculos semelhantes mostrarão que aproximadamente 94 por cento das observações caem dentro do intervalo $(\bar{x} - 2s, \bar{x} + 2s)$. Estes resultados estão certamente próximos das porcentagens teóricas de 68 e 95 para distribuições normais.

Para uma distribuição cujo histograma se pareça com o histograma da Figura 4 ou da Figura 5 não se poderia esperar encontrar para os intervalos discutidos percentagens muito aproximadas das percentagens teóricas para a distribuição normal, ainda que as percentagens freqüentemente se aproximem daqueles valores teóricos.

Pelo uso da interpretação geométrica anterior do desvio-padrão, é possível ter uma idéia grosseira do tamanho do desvio-padrão para

distribuições familiares. Consideremos, por exemplo, a distribuição de estatura para adultos do sexo masculino. Poder-se-ia supor que 95 por cento de todos estes adultos tivessem alturas compreendidas entre 5 pés e 2 polegadas e 6 pés e 2 polegadas. Como este é um intervalo de 12 polegadas e como para uma distribuição normal se espera que 95 por cento das observações caiam no intervalo $(\bar{x} - 2s, \bar{x} + 2s)$ cujo comprimento é de $4s$, pode-se admitir que $4s = 12$ polegadas ou que $s = 3$ polegadas. Uma estimativa grosseira do desvio-padrão da distribuição de freqüência dos adultos do sexo masculino é portanto de 3 polegadas. De modo análogo, poder-se-ia fazer estimativas grosseiras do tamanho do desvio-padrão para outras distribuições de freqüência familiares, e adquirir assim um sentido para o desvio-padrão como medida da variabilidade dos dados.

5 — OUTRAS MEDIDAS DESCRITIVAS

É interessante observar em reportagens jornalísticas como os diferentes grupos empregam diferentes médias para descrever a distribuição dos salários de suas indústrias. Os empregadores citam geralmente a média salarial para indicar o status econômico dos empregados. Os líderes trabalhistas, contudo, preferem usar a moda ou a mediana como indicadores do nível salarial.

A *moda* de um conjunto de medidas é definida como a medida com a freqüência máxima, se houver alguma. Assim, para o conjunto de medidas 3,3,4,4,4,5,5,6,6,7,8,9,9 a moda é 4. Em certas indústrias pode haver mais trabalhadores trabalhando na faixa de salários mais baixa que em qualquer das outras faixas e, portanto, os líderes trabalhistas preferirão, naturalmente, usar a moda, na descrição da distribuição salarial. Em muitos problemas deste tipo, devemos conhecer mais de uma medida de posição. Portanto, é provável que a moda seja de limitado valor aqui.

A *mediana* de um conjunto de medidas é definida com a medida do meio, se houver alguma, após arranjarmos as medidas em ordem de grandeza. Para o conjunto de medidas do parágrafo anterior, que está arrumado em ordem de grandeza, a mediana é 5. Se houver um número par de medidas escolhe-se a medida a meio caminho entre as duas medidas centrais. Dessa forma, se um dos 5 fosse omitido no conjunto precedente, não haveria uma medida do meio, e a mediana seria 5,5. A mediana é uma medida mais realista que a moda ou a média aritmética, para descrever o nível salarial em certas indústrias. Desde que o salário mediano seja tal que metade dos empregados receba no mínimo este tanto e metade receba no máximo este outro tanto, geralmente tem-se uma boa idéia do nível salarial pela mediana.

A média aritmética apresenta a desvantagem de que se a maior parte dos salários for notadamente baixa, mas houver uma pequena percentagem de salários muito altos, o salário médio irá indicar erroneamente um alto nível salarial. A mediana pareceria mais adequada que a média, no caso, como indicadora do que popularmente se conhece como nível salarial.

A mediana tem outra propriedade interessante. Se a variável em estudo for a renda e as rendas forem tabeladas em intervalos de 500 dólares, mas com todos aqueles que recebem mais de 20.000 dólares classificados como 20.000 dólares ou mais, não poderemos calcular a renda média, por causa das incertezas das rendas do último intervalo. A renda mediana não seria afetada por esta falta de informação e, portanto, substituiria a média aqui.

Em vista das observações precedentes sobre as propriedades atraentes da mediana, poderíamos perguntar por que os estatísticos preferem a média à mediana. A média goza de certas vantagens de cálculo sobre a mediana. Mas a principal razão pela qual se prefere a média é por ser ela uma medida muito mais útil e segura para se fazer inferências estatísticas. Como o objetivo último é resolver problemas de inferência estatística, prefere-se a média à mediana, quando se pode achar ambas as medidas.

Usa-se ocasionalmente, além do desvio-padrão, várias medidas de variação. A *amplitude total,* uma destas medidas, já foi definida na seção 4.2. É medida popular em campos como controle de qualidade industrial e metereológica. Sua popularidade baseia-se principalmente em sua facilidade de cálculo e interpretação. É simples acharmos a medida maior e a menor de um conjunto, em contraste com o cálculo do desvio-padrão do conjunto. Também é simples explicar como a amplitude mede a variação, em contraste com a explicação de como o desvio-padrão o mede. Mas é uma infeliz propriedade da amplitude total sua tendência em aumentar de tamanho à medida que aumenta o tamanho da amostra. Por exemplo, seria de se esperar que a amplitude total nos pesos de uma amostra de 10 estudantes universitários fosse consideravelmente menor que a amplitude total nos pesos de uma amostra de 100 estudantes. É possível ajustar a amplitude total para este aumento por meio de fórmulas imaginosas, mas então ela perderá a simplicidade.

As medidas de variação podem ser construídas por métodos semelhantes aos usados para definir a mediana. Ao invés de simplesmente encontrar um valor de x que divida o histograma em uma porção inferior e outra posterior, poder-se-ia achar outros dois valores de x, um separando as medidas do quarto inferior e outro separando as do quarto superior. Estas duas medidas, juntamente com a mediana, constituem os três quartis de distribuição. O quartil

menor é chamado primeiro quartil, a mediana é chamada segundo quartil e o quartil maior é chamado terceiro quartil. Uma medida simples de variação pode ser construída dos quartis, tomando-se a diferença entre o terceiro e o primeiro quartil. Esta medida, usada em alguns campos, é chamada *amplitude interquartílica.*

Poder-se-ia construir outras medidas de variação de maneira semelhante considerando decis no lugar de quartis. Decis são valores de x que dividem o histograma em décimos. Assim, poder-se-ia usar a diferença entre o nono decil e o primeiro decil, como medida de variação. As medidas deste tipo geral possuem certas vantagens sobre o desvio-padrão semelhante às vantagens da mediana sobre a média. Mas, a mesma espécie de superioridade do desvio-padrão existe aqui como existe para a média sobre a mediana, quando se trata de resolver problemas de inferência estatística.

Estas idéias podem ser ampliadas, se introduzirmos percentis, valores de x que dividem o histograma em 100 partes iguais. Os percentis são muito usados em psicologia e testes educacionais. Por exemplo, na comparação de resultados de estudantes submetidos a exames padronizados com a média nacional.

Sem dúvida, não poucos escolares já fizeram sua própria experiência, e sabem qual a sua posição num teste de aptidão escolar.

6 — DESCRIÇÃO DE DADOS AMOSTRAIS — ILUSTRAÇÕES ADICIONAIS

Esta seção destina-se à revisão de solução de problemas referentes ao material apresentado no capítulo ou, se preferirem, como ilustração adicional das técnicas de solução de problemas apresentados nas várias seções do capítulo. O número da seção aparece vinculado a cada parte do problema com várias partes, para facilitar o uso das mesmas.

Exemplo 1. Os seguintes dados já classificados representam o número de horas semanais de trabalho, com relação a 100 trabalhadores.

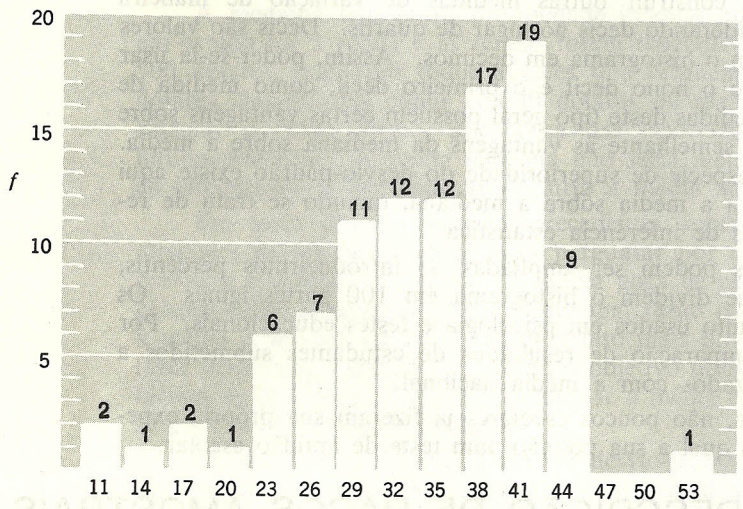
x	11	14	17	20	23	26	29	32	35	38	41	44	47	50	53
f	2	1	2	1	6	7	11	12	12	17	19	9	0	0	1

Recorrendo a estes dados: (a) 3 — faça o histograma; (b) 4.1 — calcule a média; (c) 4.2 — calcule o desvio-padrão; (d) 4.3 — calcule o percentual de medidas que cai dentro dos dois

intervalos $(\bar{x} - s, \bar{x} + s)$ e $(\bar{x} - 2s, \bar{x} + 2s)$; (e) 5 — estime a mediana examinando o histograma; (f) 5— estime o primeiro e terceiro quartis e a amplitude interquartílica, examinando o histograma.

Seguem as soluções:

(a)



(b)

x_i	f_i	$x_i f_i$
11	2	22
14	1	14
17	2	34
20	1	20
23	6	138
26	7	182
29	11	319
32	12	384
35	12	420
38	17	646
41	19	779
44	9	396
47	0	
50	0	
53	1	53
Totais	100	3.407

Usando a fórmula (4)

$$\bar{x} = \frac{3407}{100} = 34{,}07 = 34{,}1$$

(c)

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
11	2	−23,1	533,61	1067,22
14	1	−20,1	404,01	404,01
17	2	−17,1	292,41	584,82
20	1	−14,1	198,81	198,81
23	6	−11,1	123,21	739,26
26	7	− 8,1	65,61	459,27
29	11	− 5,1	26,01	286,11
32	12	− 2,1	4,41	52,92
35	12	0,9	0,81	9,72
38	17	3,9	15,21	258,57
41	19	6,9	47,61	904,59
44	9	9,9	98,01	882,09
47	0	12,9	166,41	0
50	0	15,9	252,81	0
53	1	18,9	357,21	357,21
Totais	100			6204,60

Usando a fórmula (10)

$$s = \sqrt{\frac{6204{,}60}{99}} = \sqrt{62{,}7} = 7{,}9$$

(d)

$$(\bar{x} - s, \bar{x} + s) = (26{,}2.\ 42{,}0)$$
$$(\bar{x} - 2x, \bar{x} + 2s) = (18{,}3,\ 49{,}9)$$

Do histograma em (a) será observado que os limites do intervalo no qual 26,2 cai estão entre 24,5 e 27,5. Como a distância de 26,2 a 27,5 é de 1,3 unidades e a distância de 24,5 a 27,5 é de 3 unidades, a fração 1,3/3 = 0,43 das 7 medidas dentro deste intervalo são tratados como estando à direita de 26,2. Mas 43 por cento de 7 é 3, se arredondarmos; donde 3 destas medidas deverem ser incluídas na contagem total. Cálculos semelhantes para o intervalo em que está 42,0 dá a fração 2,5/3 = 0,83. O número total de medidas dentro do intervalo (26,2, 42,0) é, portanto

$$\frac{1,3}{3}(7) + 11 + 12 + 12 + 17 + \frac{2,5}{3}(19) = 71$$

Conclusão: 71 por cento cai dentro do intervalo. De modo semelhante, o número de medidas fora do intervalo (18,3, 49,9) é aproximadamente

$$2 + 1 + \frac{2,8}{3}(2) + 1 = 6$$

Portanto, aproximadamente 94 por cento cai dentro do intervalo.

e) Cálculos refinados dão 35,5; donde 35 ou 36 ser um cálculo satisfatório.

f) Cálculos refinados dão:

Primeiro quartil = 29,1
Terceiro quartil = 40,1

Amplitude interquartílica = 40,1 − 29,1 = 11

São cálculos satisfatórios 29, 40 e 11 respectivamente.

Exemplo 2. Os seguintes dados fornecem os salários semanais de 100 trabalhadores de um país estrangeiro. (a) 2 — Classifique os dados numa tabela de freqüência. Escolha um intervalo de classe de comprimento 2 e começe 1/2 unidade abaixo da medida menor. Assim, escolha limites 38,5 — 40,5, 40,5 — 42,5 etc. e pontos de classe 39,5, 41,5 etc.; (b) 3 — Trace o histograma; (c) 4.1 — calcule a média; (d) 4.2 — Calcule o desvio-padrão; (e) 4.3 — Calcule o percentual de medidas que caem dentro do intervalo $(\bar{x} - s, \bar{x} + s)$; (f) 5 — Estime a mediana; (g) 4.1 — Calcule a média dos dados não classificados e compare com a resposta da parte (c); (h) 4.2 — Calcule o desvio-padrão dos dados não classificados e compare com a resposta da parte (d).

Usando a fórmula (4)

$$\bar{x} = \frac{5000,0}{100} = 50$$

Usando a fórmula (10)

$$s = \sqrt{\frac{1603,00}{99}} = 4,0$$

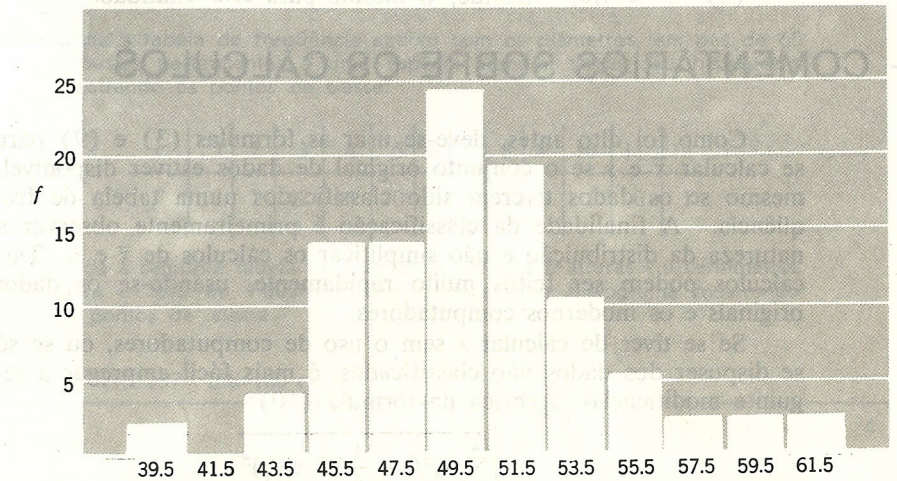
(e) $(\bar{x} - s, \bar{x} + s) = (46,54)$. Medidas dentro do intervalo total

$$\frac{0,5}{2}(14) + 14 + 25 + 21 + \frac{1,5}{2}(10)$$

49	47	51	48	50	46	53	46	45	50	49	50	50	47
56	51	46	47	54	53	48	50	51	50	60			
51	46	48	52	52	46	61	52	49	50	45	57	54	51
60	50	56	52	44	49	45	51	50	40	46			
54	47	50	55	55	47	48	53	50	49	45	50	50	51
47	54	43	53	55	50	53	52	52	51	47			
51	48	45	44	50	52	49	51	51	47	53	49	46	61
49	52	48	39	46	52	51	57	49	45	50			

(a)

x	39,5	41,5	43,5	45,5	47,5	49,5	51,5	53,5	55,5	57,5	59,5	61,5
f	2	0	3	14	14	25	21	10	5	2	2	2



(c) e (d)

x_i	f_i	$x_i f_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
39,5	2	79,0	−10,5	110,25	220,50
41,5	0	0	− 8,5	72,25	0
43,5	3	130,5	− 6,5	42,25	126,75
45,5	14	637,0	− 4,5	20,25	283,50
47,5	14	665,0	− 2,5	6,25	87,50
49,5	25	1237,5	− 0,5	0,25	6,25
51,5	21	1081,5	1,5	2,25	47,25
53,5	10	535,0	3,5	12,25	122,50
55,5	5	277,5	5,5	30,25	151,25
57,5	2	115,0	7,5	56,25	112,50
59,5	2	119,0	9,5	90,25	180,50
61,5	2	123,0	11,5	132,25	264,50
Totais	100	5000,0			1603,00

ou

$$3{,}5 + 14 + 25 + 21 + 7{,}5 = 71$$

Donde, 71 por cento.

(f) 50

(g) $\bar{\bar{x}} = 50{,}0$; Donde, o mesmo para esta exatidão.

(h) $\bar{x} = 4{,}0$; Donde, o mesmo para esta exatidão.

7 — COMENTÁRIOS SOBRE OS CÁLCULOS

Como foi dito antes, deve-se usar as fórmulas (3) e (9) para se calcular $\bar{x}$ e s se o conjunto original de dados estiver disponível, mesmo se os dados tiverem sido classificados numa tabela de freqüência. A finalidade da classificação é primeiramente observar a natureza da distribuição e não simplificar os cálculos de $\bar{x}$ e s. Tais cálculos podem ser feitos muito rapidamente, usando-se os dados originais e os modernos computadores.

Se se tiver de calcular s sem o uso de computadores, ou se só se dispuser dos dados não classificados, é mais fácil empregar a seguinte modificação algébrica da fórmula (10):

$$s = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 f_i - \frac{1}{n}(\Sigma\, x_i f_i)^2}{n-1}}$$

Esta fórmula elimina a necessidade de se calcular as diferenças $x_i - \bar{x}$; todavia, se x_i e f_i forem números grandes, a aplicação da fórmula será muito trabalhosa. E se não se cuidar da exatidão dos dígitos, o resultado final será muito pouco exato. Portanto é preciso capacidade de avaliação para se decidir se esta fórmula é melhor ou pior que a fórmula (10) para o cálculo de s. Como esta fórmula está sendo apresentada aqui a título de informação, para que os estudantes saibam que esta fórmula existe, este tipo de decisão não será necessário na solução dos problemas que seguem. Mas os estudantes poderão usar esta fórmula em vez da (10), se preferirem.

EXERCÍCIOS

Seção 1

1 — Lance uma moeda 500 vezes, registrando o número de caras obtidas em conjuntos de 50. Calcule a proporção cumulativa de caras obtidas em forma decimal e observe se o conjunto resultante de 10 proporções parece aproximar-se de 1/2.

2 — Lance um dado 500 vezes, registrando apenas o número de vezes que o número 1 ou 6 aparecer, em conjuntos de 50 lances. Calcule a proporção cumulativa de tais resultados e observe se tais proporções parecem aproximar-se de 1/3.

Seção 3

3 — Dada a tabela de freqüência abaixo com os diâmetros, em pés de 60 arbustos pertencentes a uma espécie comum, traçar seu histograma mostrando os pontos de classe.

x	1	2	3	4	5	6	7	8	9	10	11	12
f	1	7	12	16	10	4	5	2	1	1	0	1

4 — Dada a seguinte tabela de freqüência relativa a alturas em centímetros para o caso de 1 000 estudantes, traçar seu histograma mostrando os pontos de classe.

x	155–157	158–160	etc.												
f	4	10	25	52	90	145	188	180	130	90	54	26	4	1	1

5 — Os seguintes dados referem-se ao tempo de viagem de ida e volta ao trabalho em horas por dia, para um grupo de mecânicos de aviação. Traçar o histograma considerando o tempo contínuo.

menos	de 1	hora	80
1	a	2	42
2	a	3	7
3	a	4	4
4	a	5	3
5	a	6	2

6 — Dada a seguinte distribuição de freqüências do número de crianças nascidas de mulheres entre 40 e 44 anos, traçar o histograma.

x	0	1	2	3	4	5	6	7	8	9	10 ou mais
f	1230	1520	1545	962	537	301	174	108	69	51	73

7 — Que tipo de distribuição você esperaria que possuíssem as médias finais de estudantes secundários? Esboce sua idéia da natureza desta distribuição com as unidades adequadas dispostas sobre o eixo dos **x**.

Seção 4.1

8 — Para o histograma do problema 3, calcule o valor de $\bar{x}$. Faça o mesmo para os histogramas dos problemas 4 e 5.

9 — Para os dados do problema 3, calcule $\bar{x}$.

10 — Para os dados do problema 4, calcule $\bar{x}$.

Seção 4.2

11 — Sem classificar os dados, calcule (a) a média; (b) o desvio-padrão para o seguinte conjunto de pesos de 11 crianças: 39, 52, 40, 45, 46, 55, 48, 40, 43, 47, 44.

12 — Sem classificar os dados, calcule (a) a média; (b) o desvio-padrão para o seguinte conjunto de médias de 20 estudantes: 2,4, 1,2, 1,4, 2,4, 1,1, 1,8, 1,9, 1,4, 1,8, 3,2, 2,4, 2,2, 2,4, 1,8, 3,6, 1,8, 1,2, 2,4, 2,0, 3,4.

13 — Calcule **s** para os dados do problema 3.

14 — Calcule **s** para os dados do problema 4.

Seção 4.3

15 — Para o histograma do problema 3, usando os resultados dos problemas 9 e 13, calcule as porcentagens aproximadas dos dados que caem dentro dos intervalos $\bar{x} \pm s$ e $\bar{x} \pm 2s$.

16 — Para o histograma do problema 4, usando os resultados do problema 10 e 14, calcule as porcentagens aproximadas dos dados que caem dentro dos intervalos $\bar{x} \pm s$ e $\bar{x} \pm 2s$.

17 — Mostre que

(a) $$\frac{1}{n}\sum_{i=1}^{n}(x_i + c) = \bar{x} + c, \text{ (b) } \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

18 — Suponhamos que você descubra que 5 por cento de tiros ao alvo estivessem a uma distância de mais de três desvios-padrão do centro. Qual sua conclusão sobre a distribuição dos tiros?

19 — Se os tamanhos de sapatos de rapazes (ou moças) universitários fossem admitidos como possuindo uma distribuição normal, que desvio-padrão você estimaria para o tamanho do sapato se, por seu conhecimento de tamanhos de sapatos, você estimasse um intervalo de 2 desvios-padrão em relação à média?

20 — Que se pode dizer sobre a distribuição, se $s = 0$?

21 — Se as notas de um conjunto de provas são modificadas (a) pelo acréscimo de 10 pontos a todas as notas; (b) pela duplicação de todas as notas, que efeito tais alterações terão sobre a média e o desvio-padrão?

22 — Cite algum tipo de dados segundo o qual se pode sentir que o desvio-padrão tenderá a exagerar a quantidade de variação presente nos dados.

23 — Calcule o desvio-padrão para a distribuição discreta consistente de dois pontos $x = -2$ e $x = 2$ com freqüências de 50 cada um e comente sobre a interpretação dos 68 por cento e 95 por cento do **s**, aqui.

Seção 5

24 — Cite algum tipo de dados para o qual você sente que a mediana seria uma medida de posição mais adequada que a média.

25 — Ache a mediana e a amplitude total, para os dados do problema 11.

26 — Ache a mediana e a amplitude total, para os dados do problema 12.

27 — Ache a amplitude interquartílica, para os dados do problema 12.

Gerais

28 — Os dados que seguem fornecem os dados da pressão sangüínea de 50 adultos:

120, 115, 108, 111, 116, 132, 93, 107, 114, 118, 126, 124, 117, 102, 155, 123, 124, 117, 152, 130, 114, 126, 133, 120, 148, 135, 110, 100, 120, 108, 125, 136, 142, 113, 147, 155, 132, 110, 136, 125, 96, 112, 120, 118, 115, 122, 168, 160, 126, 154.

(a) Disponha os dados numa tabela de freqüência. Escolha um intervalo de classe de comprimento 5 e comece 1/2 unidade abaixo da medida menor.

(b) Trace o histograma para a distribuição da freqüência em (a).

(c) Examinando o histograma de (b), calcule o valor da média.

(d) Calcule a média dos dados não classificados.

(e) Calcule a média dos dados classificados.

(f) Examinando o histograma de (b), calcule o valor do desvio-padrão.

(g) Calcule o desvio-padrão dos dados não classificados.

(h) Calcule o desvio-padrão dos dados classificados.

(i) Examinando o histograma de (b), calcule o valor dos três quartis.

29 — A rentabilidade de um conjunto escolhido de ações da Bolsa de Nova York aparece a seguir:

14,6	25,4	12,2	9,3	20,3	7,6
13,4	9,6	14,7	25,3	13,7	10,1
11,9	18,9	8,5	12,3	14,8	8,6
28,2	16,2	12,8	16,2	12,1	14,0
15,4	27,5	30,4	16,2	16,5	11,6
9,7	21,1	12,9	9,8	30,4	13,3
11,7	8,7	31,6	19,5	22,4	8,3

(a) Disponha os dados numa tabela de freqüência. Escolha um intervalo de classe de 2 de comprimento e comece 1/2 unidade abaixo da medida menor. Prossiga perfazendo os passos de (b) e (i) do problema 28.

LEMBRETE AOS INSTRUTORES

A Tabela XII do apêndice é um conjunto de dados obtidos num estudo para determinar a relação entre o sucesso do estudante no curso e sua avaliação da eficiência do professor. Dispõe-se de várias variáveis para estudo. O instrutor pode escolher uma destas variáveis e tantos dados quantos desejar para elaborar outros problemas. Tais dados, em conjunto, destinam-se a problemas para as classes que tenham acesso a computadores digitais. São também úteis na elaboração de problemas para os últimos capítulos.

3

PROBABILIDADES

1 — INTRODUÇÃC

Como ficou dito no Capítulo 1, as soluções para os problemas estatísticos nele propostos são dadas em termos de afirmações de probabilidade. Embora a probabilidade se aplique a grande número de situações práticas, a compreensão daquilo em que ela consiste se torna muito mais simples se aplicada a situações não práticas, tais como as que ocorrem em certos jogos de azar. Por esta razão a definição e as regras de probabilidade são apresentadas em termos de situações ideais, mas admite-se que as mesmas regras possam depois ser aplicadas a problemas estatísticos práticos.

Antes de discutirmos a probabilidade, é necessário discutirmos experimentos que possam ser repetidos ou possam ser concebidos como susceptíveis de repetição. O lançamento de uma moeda, a leitura da temperatura diária num termômetro ou a contagem do número de ovos estragados de uma cesta são exemplos de experimentos repetitivos simples. Um experimento no qual vários coelhos são alimentados com diferentes rações na tentativa de se determinarem as propriedades do crescimento relativo possuídas por essas rações, só pode ser feito uma vez com estes animais. Mas podemos considerar o experimento como o primeiro de uma série ilimitada de experimentos semelhantes e, portanto, pode ser considerado como repetitivo. A escolha de uma amostra de uma população é um experimento repetitivo e é, certamente, o tipo de experimento que possui interesse particular na solução de problemas estatísticos.

Consideremos um experimento repetitivo simples, como, por exemplo, o lançamento de uma moeda duas vezes ou, o que é equi-

valente, o lançamento de duas moedas, ao mesmo tempo. Nesse experimento há 4 ocorrências possíveis; são elas:

Ca Ca, Ca C, C Ca, CC.

O símbolo CaC, por exemplo, significa que é obtida uma cara no primeiro lançamento e uma coroa no segundo. Se o experimento tivesse consistido no lançamento da moeda três vezes, haveria oito ocorrências possíveis, que seriam as seguintes:

CaCaCa, CaCaC, CaCCa, CCaCa, CaCC, CCaC, CCCa, CCC.

As três letras de cada grupo exprimem as ocorrências dos três lançamentos na ordem em que foram realizados. Um experimento como a leitura da temperatura de um termômetro, todavia, tem um número infinito de possíveis ocorrências, visto que a temperatura é um tipo de variável contínua. Nesta discussão de probabilidades apenas os experimentos com um número finito de ocorrências possíveis são considerados. Outros tipos serão discutidos no Capítulo 4.

Para qualquer experimento ao qual a probabilidade vai ser aplicada, é necessário fazer a lista de todas as ocorrências possíveis. Esta lista deve ser tal que, quando se realiza o experimento, ocorra exatamente um dos eventos. No experimento do lançamento da moeda três vezes, o interesse estava centrado sobre o fato de a moeda mostrar cara ou coroa, em cada um dos lançamentos. Portanto, todas as possíveis ocorrências foram listadas antes. Ao selecionarmos um dígito de uma tabela aleatória (Tabela XI) no apêndice, poderíamos estar interessados em saber qual o dígito obtido, caso em que há 10 ocorrências possíveis que correspondem a 0,1, ..., 9. Contudo, poderíamos estar interessados apenas em saber se o dígito era menor do que 3. Neste caso, só haveria duas ocorrências possíveis do experimento, isto é, obter um dígito menor do que 3 ou obter um dígito pelo menos igual a 3. Um experimento de jogo de azar que usaremos freqüentemente com propósitos ilustrativos é o de retirarmos uma bola de uma caixa que contém bolas de várias cores. Suponhamos, por exemplo, que uma caixa contenha 3 bolas vermelhas, 2 pretas e 1 verde. O interesse está em saber qual a cor da bola retirada e não no tipo particular da bola. Aqui há 3 ocorrências possíveis correspondentes às 3 cores. Outro experimento interessante, referente a jogos de azar é o de jogarmos dois dados. Suposto que se possa distinguir entre os dois dados bem como saber o número de pontos mostrados por cada dado, haverá, então, 36 ocorrências possíveis, pois cada dado tem 6 possíveis ocorrências, e tais ocorrências poderão surgir aos pares, em todas as combinações possíveis.

QUADRO 1

11	21	31	41	51	61
12	22	32	42	52	62
13	23	33	43	53	63
14	24	34	44	54	64
15	25	35	45	55	65
16	26	36	46	56	66

O primeiro número de cada par denota o número que surgiu em um dos dados e o segundo denota o número que apareceu no outro. Se os dois dados não puderem ser diferenciados, será preciso jogá-los separadamente e não simultaneamente.

2 — ESPAÇO AMOSTRAL

Ao desenvolvermos a teoria da probabilidade, será conveniente visualizarmos as coisas geometricamente e representar cada uma das possíveis ocorrências de um experimento, por meio de um ponto. Usaríamos, assim, 8 pontos, no lançamento tríplice da moeda. Não faz diferença que pontos serão escolhidos quando se sabe que ponto corresponde à possível ocorrência. Cada ponto está marcado com uma letra ou símbolo para indicar que ocorrência ele representa. Como os sucessos possíveis são, freqüentemente, chamados eventos simples, a letra e com um índice correspondente ao número de ocorrências da lista de ocorrências possíveis é geralmente usada na marcação do ponto. No experimento do lançamento da moeda, por exemplo, poderíamos marcar os 8 pontos por meio de e_1, e_2, ... e_8. Assim e representaria o evento CaCaCa e e_2 o evento CaCaC. Como um símbolo como CaCaC é suficientemente claro, não há necessidade de símbolos adicionais. Todavia é mais fácil escrevermos e_2 do que CaCaC e, neste sentido, o símbolo e possui uma vantagem sobre o outro. Ao desenvolvermos as regras de probabilidades, será conveniente darmos nomes a estes conjuntos básicos de pontos. Formalmente isto é feito assim:

Definição. *O conjunto de pontos que representa as ocorrências possíveis de um experimento chama-se espaço amostral do experimento.*

A Figura 1 mostra o espaço amostral do lançamento das moedas onde a indicação dos pontos são símbolos que aparecem diretamente sobre eles.

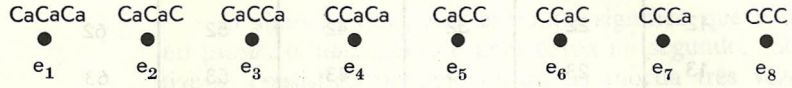


FIGURA 1. Espaço amostral para o experimento do lançamento de uma moeda.

Um espaço amostral natural a ser escolhido para o experimento de selecionarmos um dígito da tabela de dígitos aleatórios consiste em 10 pontos sobre o eixo do x correspondentes aos inteiros 0,1,... 9. Isto aparece na Figura 2. A letra e não foi usada na Figura 2 porque a marcação de um ponto por meio de suas coordenadas é tão simples quanto se pode desejar.

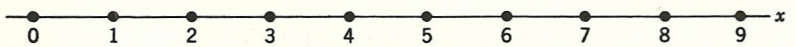


FIGURA 2. Espaço amostral para o experimento de dígitos aleatórios.

Para o experimento relativo à escolha da bola colorida da caixa de bolas, o espaço amostral consiste em 3 pontos marcados na ordem das cores dadas e que aparecem na Figura 3.

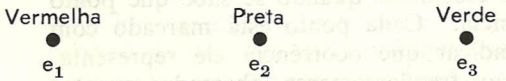


FIGURA 3. Espaço amostral para o experimento das bolas coloridas.

O espaço amostral conveniente para o experimento do lançamento dos 2 dados é um quadro duplo de 36 pontos com 6 linhas e 6 colunas, ligados aos elementos do Quadro 1. Este espaço amostral não será mostrado aqui, porque basta o Quadro 1 e sua imaginação. Os símbolos do Quadro 1 são altamente descritivos das ocorrências experimentais e são também muito simples: não há necessidade de se introduzir um novo conjunto de símbolos aqui.

3 — PROBABILIDADE DE UM EVENTO

Até aqui não houve razão aparente de introduzirmos uma representação geométrica relativa a ocorrências de um experimento;

todavia, com o desenvolvimento posterior da teoria, torna-se evidente a vantagem de tal método.

O próximo passo na construção de um modelo matemático para o experimento é vincular números aos pontos do espaço amostral, que representarão as freqüências relativas, isto é, a proporção de vezes que se espera que tais eventos ocorram. Se o experimento de se lançar uma moeda três vezes fosse repetido um grande número de vezes e se se conservasse um registro cumulativo da proporção destes experimentos que produziram, digamos, 3 caras, poder-se-ia esperar que a proporção se aproximasse de 1/8, porque cada uma das oito ocorrências possíveis deveria ocorrer o mesmo número aproximado de vezes. Experimentos reais desta espécie geralmente mostram que tais expectativas são justificadas, se a moeda não estiver viciada e for bem lançada. Em vista de tais considerações, o número 1/8 deveria estar ligado a cada ponto do espaço amostral mostrado na Figura 1. O número ligado ao ponto marcado com e_i no espaço amostral chama-se probabilidade do evento e_i e é simbolizado por $P[\ e_i\ ]$. Assim, no experimento do lançamento das moedas, cada um dos eventos $e_1, e_2, \ldots, e_8$ possui a probabilidade 1/8.

Se efetuarmos muitas vezes o experimento da seleção de dígitos de uma tabela aleatória descobriremos que cada um dos dez dígitos 0, 1, ..., 9 será obtido com aproximadamente a mesma freqüência relativa e, portanto, a freqüência relativa experimental para cada um dos dígitos se aproximará de 1/10. Com base em tal experiência dever-se-ia apontar a probabilidade de 1/10 para cada um dos pontos do espaço amostral da Figura 2.

O experimento do lançamento dos dois dados é tratado do mesmo modo que o do lançamento da moeda. A simetria e a experiência sugerem que se deve atribuir a probabilidade de 1/36 a cada ponto do espaço amostral correspondente ao Quadro 1. A situação do experimento correspondente ao espaço amostral da Figura 3 é um pouco diferente dos precedentes. Não é mais verdade que se deva esperar a ocorrência de cada sucesso, com a mesma freqüência relativa. Se as bolas forem bem misturadas antes de cada retirada e a bola retirada for reposta na caixa de modo que se mantenha inalterada sua composição, deveríamos esperar que a bola preta aparecesse duas vezes mais que a verde e a vermelha três vezes mais que a verde. Neste experimento amostral repetido deveríamos esperar que as freqüências relativas das três cores (vermelho, preto e verde) se aproximassem de 3/6, 2/6 e 1/6 respectivamente. Assim as probabilidades dos 3 pontos e_1, e_2 e e_3 da Figura 3 seriam 3/6, 2/6 e 1/6, respectivamente.

Os experimentos anteriores ilustram como se procede, em geral, na designação de pontos para um espaço amostral. Se o experimento for do tipo para o qual a simetria e considerações semelhantes sugerem quais as freqüências relativas que devem ser esperadas para os vários sucessos, tais freqüências relativas são as probabilidades dos pontos correspondentes. Esta foi a base para acharmos as probabilidades de cada ponto amostral no caso do lançamento de dados, do jogo da moeda e da retirada das bolas. Se não dispusermos de tais considerações de simetria (ou de analogia) mas pudermos fazer experiências com o problema em questão, os dados obtidos de tais experiências nos darão a pista para acharmos as probabilidades. Os dados que compõem o espaço amostral da Figura 2 foram obtidos, em parte, da experiência e, em parte, da fé nos indivíduos que construíram a Tabela XI. Há vários métodos de construção de tabelas de dígitos aleatórios, alguns muito complicados. Em todas estas tabelas, deve-se esperar que cada dígito ocorra mais ou menos o mesmo número de vezes e que não haja padrões discerníveis nas seqüências dos dígitos. Mas, como tais conjuntos de dígitos estão baseados em artifícios físicos que se supõe produzam dígitos com tais propriedades, não é razoável esperar que tais dígitos se comportem de maneira ideal. Tudo que se pode esperar é uma boa aproximação.

Como as probabilidades dos pontos de um espaço amostral ou são freqüências relativas esperadas baseadas em considerações de analogia ou freqüências relativas baseadas em uma longa série de experiências, as probabilidades são números de 0 a 1 e sua soma deve ser 1, porque a soma de um conjunto completo de freqüências relativas é sempre 1. Nos experimentos relativos à moeda, bolas e dados, as probabilidades somam 1 porque foram construídas deste modo. Se as probabilidades com experimentos referentes a dígitos aleatórios forem totalmente baseadas em freqüências relativas obtidas ao longo de prolongadas experiências, tais probabilidades somarão 1, porque a soma de todas as freqüências relativas é 1.

Ora, em qualquer situação experimental, real ou acadêmica, é privilégio do estatístico atribuir às ocorrências do experimento as probabilidades que ele deseja, contanto que sejam números entre 0 e 1 e somem 1. Naturalmente se ele for sensível, procurará atribuir números que representem o que julga ou sabe ser a freqüência relativa destes sucessos; de outro modo seu modelo matemático provavelmente não representaria a realidade de modo satisfatório, e, portanto, as conclusões obtidas com base no modelo provavelmente estariam erradas.

É geralmente muito fácil atribuirmos freqüências relativas satisfatórias aos casos possíveis dos jogos de azar; contudo, não é este o caso da maioria das experiências da vida real. Por exemplo, se a experiência consiste em selecionar, ao acaso, um indivíduo da população de uma cidade para se saber se ele morrerá no ano seguinte, então para se atribuir aqui uma freqüência relativa esperada não há maneira mais satisfatória que usar a experiência das companhias de seguros. Se alguém estivesse interessado na determinação de prêmios de seguros convenientes, seria necessário atribuir à mortalidade, nas várias idades, freqüências relativas esperadas. Estas são geralmente escolhidas a partir dos valores obtidos da experiência das companhias de seguros, ao longo dos anos. Já que os índices de mortalidade têm decrescido através dos anos para a maioria dos grupos etários, qualquer tabela de mortalidade baseada na experiência passada provavelmente estará desatualizada na predição do futuro. Assim, as freqüências relativas fixadas à base de experiências do passado podem não ser muito aproximadas das freqüências relativas reais existentes atualmente, e conseqüentemente as probabilidades calculadas a partir delas não serão muito precisas. Felizmente para as companhias de seguros, os prêmios calculados à base de experiências passadas são maiores do que seriam se tivessem sido baseados em experiências mais atualizadas.

Em vista da discussão anterior, segue-se que a probabilidade de um acontecimento deve ser interpretada como uma freqüência relativa teórica ou ideal deste acontecimento. Não implica isso que observada freqüência relativa do acontecimento deva necessariamente aproximar-se da probabilidade deste acontecimento para um número de experiências progressivamente crescente, porque se pode ter escolhido um modelo incorreto; contudo espera-se que isto ocorra. Assim, se alguém tem dois dados que se supõe não viciados, poderia esperar que a freqüência relativa observada de um total de 4 pontos fosse aproximar-se da probabilidade 1/6 com um número progressivamente crescente de lançamentos; mas não ficaria muito surpreendido se ela não se aproximasse de 1/6 devido a imperfeições relativas a qualquer artigo manufaturado e a dificuldade de se simular um experimento ideal. Neste particular, deve-se notar que os operadores de casas de jogo fazem bem, financeiramente, por admitir que os dados se comportem conforme é esperado. Eles certamente jogaram dados o bastante para pôr à prova tal suposição. É claro que se a experiência mostra que um dado não se comporta conforme o esperado, eles o substituem rapidamente por um novo dado.

A construção de modelos teóricos para explicar a natureza é a principal função dos cientistas. Se tais modelos forem realistas, as conclusões deles obtidas serão, provavelmente, realistas. Um modelo probabilístico é um modelo destinado a permitir que se tirem conclusões sobre as freqüências relativas de ocorrências experimentais.

Embora a probabilidade esteja sendo interpretada aqui em termos de freqüência relativa esperada, também é usada por alguns indivíduos como medida do grau da crença na ocorrência futura de um evento, mesmo não havendo conceitualmente experimento repetível. Assim, por exemplo, poderiam estar querendo atribuir certa probabilidade ao evento "não haverá guerra mundial nos próximos 10 anos". Tal tipo de probabilidade corresponderia às vantagens de apostas que o indivíduo daria à não ocorrência da guerra. Tais tipos de probabilidades, chamadas pessoais ou subjetivas, são freqüentemente usadas nos negócios e serão discutidas no Capítulo 14.

4 — PROBABILIDADE DE UM EVENTO COMPOSTO

Agora que construímos um modelo geométrico para um experimento que consiste num conjunto de pontos marcados com as letras e_1, e_2, ... para representar todas as ocorrências possíveis e um conjunto associado de correspondentes probabilidades $P\{e_1\}$, $P\{e_2\}$, ..., chegou o tempo de discutirmos a probabilidade de eventos compostos. As ocorrências possíveis de um espaço amostral e_1, e_2, ... chamam-se *eventos simples*. O *evento composto* é definido como uma coleção de eventos simples. Por exemplo, o evento de obter exatamente duas caras na experiência do lançamento da moeda da Figura 1 é um evento composto que consiste em eventos simples e_2, e_3 e e_4. Do mesmo modo, o evento de obter um dígito aleatório menor que 4 no experimento do dígito aleatório da Figura 2 é um evento composto que consiste em 4 eventos simples $x = 0, 1, 2, 3$. Eventos compostos são geralmente simbolizados pelas letras maiúsculas A ou B ou C.

Ora, como as probabilidades atribuídas aos eventos simples da Figura 1, a saber: 1/8, representam as freqüências relativas esperadas para sua ocorrência, esperar-se-ia que o evento composto consistente nos eventos simples e_2, e_3 e e_4 ocorressem em cerca de 3/8 de tais experimentos se fossem realizados em grande número. Analogamente, para o experimento representado pela Figura 2 esperar-se-ia que o evento composto consistente nos eventos simples $x = 0, 1, 2, 3$ ocorresse em cerca de 4/10 de tais experimentos, se realizados bastante vezes. Em vista de tais expectativas e porque a probabilidade foi introduzida como idealização da freqüência relativa, a definição de probabilidade que damos a seguir, referente a evento composto, deverá parecer bastante razoável.

(1) Definição. *A probabilidade de que ocorra um evento* A *composto é a soma das probabilidades dos eventos simples dos quais se compõe.*

Como ilustração, se A for o evento de obter duas caras no lançamento triplo de uma moeda, segue-se da definição e do espaço amostral da Figura 1 que

$$P\{A\} = P\{e_2\} + P\{e_3\} + P\{e_4\} = \tfrac{3}{8}$$

De modo análogo, se B for o evento de obter um dígito menor que 4, ao selecionarmos dígitos aleatórios, segue-se da definição e da Figura 2 que

$$P\{B\} = P\{0\} + P\{1\} + P\{2\} + P\{3\} = \tfrac{4}{10}$$

Vejamos mais um exemplo. Se C simbolizar o evento de obtermos uma bola vermelha ou verde para o experimento cujo espaço amostral é expresso pela Figura 3, como C é composto de e_1 e e_3, segue-se que

$$P\{C\} = P\{e_1\} + P\{e_3\} = \tfrac{4}{6}$$

Como ilustração final para os quais os eventos compostos não são tão óbvios, consideremos uma vez mais a experiência do lançamento de dois dados. O Quadro 1, com pontos associados a cada ocorrência e com a probabilidade 1/36 ligada à cada ponto, pode servir como espaço amostral. Primeiro, seja E o evento de obtermos um total de 7 pontos, nos dois dados. Os eventos simples que fornecem um total de 7 pontos são os seguintes: 16, 25, 34, 43, 52, 61. A soma das 6 probabilidades correspondentes de 1/36 cada uma dá, portanto, $P\{E\} = \frac{6}{36} = \frac{1}{6}$. Segundo, seja F o evento de obtermos um número total de pontos que seja par. Eventos simples como 11, 13, 22 etc. satisfazem à exigência de obtenção de um número total par. A soma de dois dígitos pares ou a soma de dois dígitos ímpares resulta sempre em um número par. No Quadro 1 pode-se observar que há 18 pontos deste tipo. Segue-se que $P\{F\} = \frac{18}{36} = \frac{1}{2}$. Finalmente, seja G o evento segundo o qual ambos os dados mostram, pelo menos, 4 pontos. Aqui são suficientes os eventos simples como 44, 45, 56 etc. O Quadro 1 mostra que há 9; portanto, $P\{G\} = \frac{9}{36} = \frac{1}{4}$. Em muitos experimentos de jogos de azar espera-se que os vários sucessos possíveis ocorram com a mesma freqüência relativa; por isso atribui-se a mesma probabilidade a todos os pontos do espaço amostral, no tocante a estes experimentos, isto é $1/n$, onde n significa o número total de pontos

do espaço amostral. Isto é verdade, por exemplo, nos experimentos do lançamento da moeda, escolha de dígitos aleatórios e lançamento de dados. Não o é, contudo, para o experimento da retirada de bolas coloridas. Quando o experimento for do tipo simples, isto é, quando todas as probabilidades do evento simples forem iguais, o cálculo da probabilidade de um evento composto é muito simples. Consiste simplesmente em somar a probabilidade $1/n$ tantas vezes quantos forem os eventos simples que compõem o evento composto. Assim, se o evento composto A consistir num total de $n(A)$ eventos simples, o valor de $P\{A\}$ pode ser expresso pela simples fórmula

$$P\{A\} = \frac{n(A)}{n} \tag{2}$$

No experimento do lance dos dois dados, por exemplo, a probabilidade de se obter um total de 7 pontos é conseguida pela contagem do número de pontos do espaço amostral do Quadro 1 que soma um total 7, isto é, 6 pontos, e dividindo este número pelo número total de pontos, isto é, 36.

Embora muitas vezes na vida real não seja possível usar a fórmula (2), é mais fácil trabalhar com ela do que com a fórmula da definição geral (1) que envolve a adição de probabilidades; por isso ela só é usada nas próximas seções para derivar fórmulas básicas. Pode-se mostrar que as fórmulas obtidas dessa maneira se comportam igualmente bem em relação à definição mais geral, e por conseguinte são aplicáveis a todos os tipos de problemas. Desde que somente as fórmulas são necessárias aos problemas aplicados, não haverá perda apreciável de conhecimento na resolução dos problemas práticos seguindo-se esse processo.

5 — REGRA DA ADIÇÃO

As aplicações de probabilidade freqüentemente se prendem mais a certo número de acontecimentos relacionados do que apenas a um acontecimento. Por simplicidade, consideremos dois acontecimentos, A_1 e A_2, associados a uma experiência. Alguém poderá estar interessado em saber se ambos, A_1 e A_2, ocorrerão quando a experiência for realizada. Este acontecimento conjunto é indicado pelo símbolo de produto (A_1 e A_2) e sua probabilidade por $P\{A_1 \text{ e } A_2\}$. Por outro lado, pode-se estar interessado em saber se pelo menos um dos acontecimentos A_1 e A_2 ocorrerá quando se realizar a citada experiência. Este evento é simbolizado por (A_1 ou A_2) e sua probabilidade por $P\{A_1 \text{ ou } A_2\}$. Pelo menos um dos dois eventos ocorrerá se A_1 ocorrer, mas A_2 não, se A_2 ocorrer, mas A_1 não ou se ambos A_1 e A_2 ocorrerem. Assim, a palavra "ou", aqui, significa "ou" no sentido de *ou um, ou outro, ou ambos*. O propósito desta seção é obter uma fórmula para $P\{A_1 \text{ ou } A_2\}$.

Se dois acontecimentos A_1 e A_2 possuem a propriedade de que a ocorrência de um impeça a ocorrência do outro, eles são chamados acontecimentos *mutuamente exclusivos*. Por exemplo, seja A_1 o acontecimento de obtenção de um total de 7 pontos no lançamento de dois dados e A_2 o acontecimento de obtenção de um total de 11: então, A_1 e A_2 são acontecimentos mutuamente exclusivos. Para os acontecimentos mutuamente exclusivos não há casos favoráveis a ocorrências de ambos A_1 e A_2; por conseguinte os dois eventos não possuem qualquer ponto em comum no espaço amostral. Isto aparece esquematicamente na Figura 4.

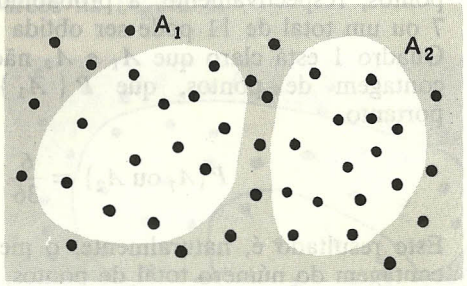


FIGURA 4. Espaço amostral com dois eventos mutuamente exclusivos.

Neste diagrama, os pontos dentro das regiões brancas marcadas com A_1 e A_2 representam os eventos simples que dão os eventos compostos A_1 e A_2 respectivamente. Se $n(A_1)$ significa o número de pontos dentro da região marcada com A_1 e $n(A_2)$ o número dentro da região marcada com A_2, então o número total de pontos associados à ocorrência ou de A_1 ou de A_2 é a soma desses números; conseqüentemente se n denotar o número total de pontos amostrais, segue-se da fórmula (2) que

$$P\{A_1 \text{ ou } A_2\} = \frac{n(A_1) + n(A_2)}{n}$$

$$= \frac{n(A_1)}{n} + \frac{n(A_2)}{n}$$

Como as duas últimas frações são precisamente as que definem $P\{A_1\}$ e $P\{A_2\}$, este resultado fornece a fórmula de adição desejada, que pode ser expressa da seguinte maneira:

(3) Regra da adição. *Quando A_1 e A_2 são eventos mutuamente exclusivos,*

$$P\{A_1 \text{ ou } A_2\} = P\{A_1\} + P\{A_2\}$$

Para mais de dois acontecimentos mutuamente exclusivos basta aplicar essa fórmula tantas vezes quanto necessário. Pode-se obter uma fórmula um pouco mais complicada para eventos não mutuamente exclusivos. Mas não haverá oportunidade de usarmos tal fórmula nas seções posteriores e a omitiremos aqui. Aparecerá num dos exercícios do fim deste capítulo.

Exemplo 1. Na ilustração precedente do lançamento dos dados, onde A_1 e A_2 significam os eventos de obtermos um total de 7 e 11 pontos, respectivamente, a probabilidade de obtermos um total de 7 ou um total de 11 pode ser obtida por meio da fórmula (3). No Quadro 1 está claro que A_1 e A_2 não têm pontos em comum e, na contagem de pontos, que $P\{A_1\} = 6/36$ e $P\{A_2\} = 2/36$; portanto

$$P\{A_1 \text{ ou } A_2\} = \frac{6}{36} + \frac{2}{36} = \frac{8}{36}$$

Este resultado é, naturalmente, o mesmo obtido anteriormente pela contagem do número total de pontos, isto é, 8, que fornece o evento composto (A_1 ou A_2) e que aplica a fórmula (2) diretamente.

Exemplo 2. Ainda como ilustração, qual é a probabilidade de se obter um total de, pelo menos, 10 pontos, no lançamento de 2 dados?

Façamos A_1, A_2 e A_3 serem os eventos de se obter um total de exatamente 10 pontos, 11 pontos e 12 pontos, respectivamente. Do Quadro 1 aparece claro que tais eventos não têm ponto em comum e que suas probabilidades são dadas por $P\{A_1\} = \frac{3}{36}$, $P\{A_2\} = \frac{2}{36}$ e $P\{A_3\} = \frac{1}{36}$. Portanto, pela fórmula (3), a probabilidade de que pelo menos um destes eventos mutuamente exclusivos ocorra é dada por

$$P\{A_1 \text{ ou } A_2 \text{ ou } A_3\} = \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{1}{6}$$

Este resultado também poderia ter sido obtido do Quadro 1 pela contagem dos casos favoráveis e totais. Embora a fórmula aqui não pareça possuir nenhuma vantagem sobre a contagem direta para problemas relacionados com o Quadro 1, é uma fórmula muito útil para problemas nos quais as probabilidades dos acontecimentos são disponíveis, mas não o são as tabelas das ocorrências possíveis.

6 — REGRA DA MULTIPLICAÇÃO

O objetivo desta seção é obter uma fórmula para $P\{A_1 \text{ e } A_2\}$ em termos de probabilidades de eventos isolados. Para fazê-lo é necessário que introduzamos a noção de probabilidade condicional. Suponhamos que se está interessado em saber se A_2 ocorrerá, sujeito à hipótese de que A_1 tenha ocorrido ou venha a ocorrer com certeza. Supõe-se aqui que A_1 e A_2 não são eventos mutuamente exclusivos. A geometria deste problema aparece na Figura 5.

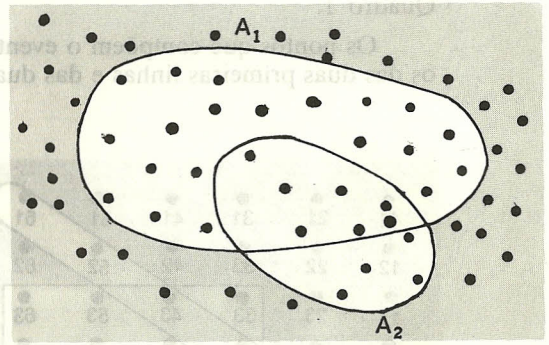


FIGURA 5. Espaço amostral para probabilidade condicional.

Como A_1 deve ocorrer, as únicas ocorrências experimentais que precisam ser consideradas são as que correspondem à ocorrência de A_1. O espaço amostral para este problema reduz-se, portanto, a eventos simples que compõem A_1. São representados na Figura 5 pelos pontos que caem dentro da região marcada A_1. Entre esses pontos, aqueles que também caem dentro da região A_2 correspondem a ocorrências de ambos A_1 e A_2. São pontos que caem nas porções de A_1 e A_2 que se sobrepõem. Façamos $n(A_1)$ significar o número de pontos de dentro de A_1, e $n(A_1 \text{ e } A_2)$ significar o número que cai dentro tanto de A_1 como de A_2. Então, sabemos pela fórmula (2) que a probabilidade de ocorrência de A_2, se o espaço amostral se restringir ao conjunto de pontos dentro de A_1,

é dado pelo quociente de $n(A_1 \text{ e } A_2) / n(A_1)$. Mas esta probabilidade é a que é significada pela probabilidade de ocorrer A_2, sujeita à condição de ocorrência de A_1.

Se esta última probabilidade for expressa pelo novo símbolo $P\{A_2|A_1\}$, teremos

$$P\{A_2|A_1\} = \frac{n(A_1 \text{ e } A_2)}{n(A_1)} \tag{4}$$

Exemplo. Como ilustração da aplicação desta fórmula, será feito um cálculo da probabilidade de que a soma de pontos obtidos no lançamento de dois dados é 7, se se sabe que os dados ostentam, pelo menos, 3 pontos cada um.

Façamos A_1 significar o evento de que os dois dados mostram pelo menos 3 pontos cada um e A_2 significar o evento de que os dois dados apresentam um total de 7 pontos. O espaço amostral deste problema aparece na Figura 6; foi obtido diretamente do Quadro 1.

Os pontos que compõem o evento A_1 são todos os pontos menos os das duas primeiras linhas e das duas primeiras colunas da Figura 6.

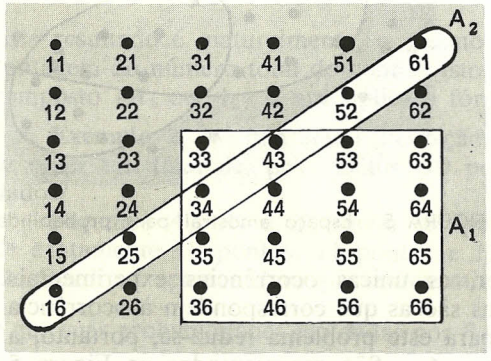


FIGURA 6. Espaço amostral para um problema de probabilidade condicional.

Aparecem dentro do retângulo branco da Figura 6. Aqui $n(A_1) = 16$. Os pontos que compõem o evento A_2 são os pontos da diagonal branca mostrados na Figura 6. O número de pontos que cai dentro de A_2 e que também cai dentro de A_1 é $n(A_1 \text{ e } A_2) = 2$. Como resultado, a fórmula (4) dá o resultado

$$P\{A_2|A_1\} = \frac{2}{16} = \frac{1}{8}$$

O que isto significa, em sentido experimental, é que no lançamento de dois dados descartamos todas essas ocorrências experimentais nas quais qualquer dos dados mostre um número de pontos menor que 3. Então, entre as ocorrências experimentais que sobram calcula-se a proporção delas que fornece um total de 7 pontos. Tal proporção, com a repetição dos experimentos, deve aproximar-se de 1/8. É interessante notar que as oportunidades de se obter um total de 7 pontos é menor, quando se sabe que ambos os dados mostram pelo menos 3 pontos, do que nos lances ordinários.

Consideremos agora a fórmula (4), mais uma vez em termos gerais. Da Figura 5 e da fórmula (2) fica claro que

$$P\{A_1\} = \frac{n(A_1)}{n}$$

e que

$$P\{A_1 \text{ e } A_2\} = \frac{n(A_1 \text{ e } A_2)}{n}$$

Dividindo a segunda destas duas expressões pela primeira e cancelando o n dará

$$\frac{P\{A_1 \text{ e } A_2\}}{P\{A_1\}} = \frac{n(A_1 \text{ e } A_2)}{n(A_1)}$$

Este resultado juntamente com a fórmula (4) fornecerá a fórmula

$$P\{A_2|A_1\} = \frac{P\{A_1 \text{ e } A_2\}}{P\{A_1\}} \tag{5}$$

Esta fórmula define a *probabilidade condicional* de A_2 dado A_1 e quando escrita em forma de produto fornece a fórmula de multiplicação fundamental para probabilidades, que pode ser expressa da seguinte maneira:

(6) Regra de multiplicação. $P\{A_1 \text{ e } A_2\} = P\{A_1\}P\{A_2|A_1\}$

Em palavras, esta fórmula estabelece que a probabilidade de que ambos os acontecimentos ocorram é igual à probabilidade de que o primeiro ocorra, multiplicada pela probabilidade condicional de que o segundo ocorra quando se sabe que o primeiro acontecimento

deve ocorrer certamente. Qualquer dos dois eventos pode ser chamado de primeiro, já que é apenas uma linguagem conveniente para discuti-los e não há ordem de tempo implicada na maneira como eles ocorrem. Embora não interesse a ordem cronológica para os dois acontecimentos A_1 e A_2 no símbolo $P\{A_2|A_1\}$, é costume chamar-se esta probabilidade condicional "a probabilidade de que A_2 vá ocorrer quando se sabe que A_1 já ocorreu". Assim, se você receber uma mão de cinco cartas de pôquer alguém poderá perguntar: "Qual é a probabilidade de que sua mão vá conter o ás de espadas se se sabe que você recebeu o ás de copas?"

Trata-se apenas de uma linguagem conveniente para discutirmos as probabilidades de uma mão de pôquer que deva conter o ás de copas, não implicando que se o ás de espadas está na mão tenha sido obtido após a obtenção do ás de copas. Para muitos pares de acontecimentos, todavia, há uma relação de ordem cronológica definida. Por exemplo, se A_1 é o acontecimento de que um concluinte de secundário vá para a faculdade e A_2 é o acontecimento de que ele vá concluir a faculdade, A_1 deve preceder A_2 no tempo.

Exemplo 1. Como ilustração da regra de multiplicação, far-se-á um cálculo da probabilidade de se obter duas bolas vermelhas na retirada de duas bolas de uma caixa que contenha três bolas vermelhas, duas pretas e uma verde. Suporemos aqui que a primeira bola retirada não será devolvida à caixa antes que seja feita a segunda retirada.

Façamos A_1 significar o evento da retirada de uma bola vermelha na primeira tentativa e A_2 o evento pelo qual obtemos uma bola vermelha na segunda tentativa. Para podermos continuar usando a fórmula (2) ao invés da fórmula mais geral da definição (1), será necessário atribuirmos um número a cada bola e usarmos os 6 pontos do espaço amostral que representa a primeira retirada, como mostra a Figura 7. Os primeiros três pontos representarão bolas vermelhas, os outros dois, bolas pretas e finalmente o último, bola verde.

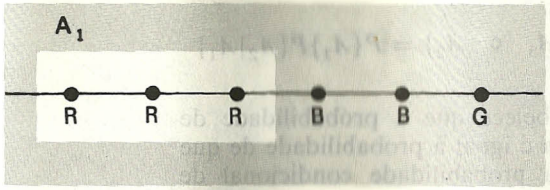


FIGURA 7. Espaço amostral da primeira retirada.

Então, pela fórmula (2),

$$P\{A_1\} = \frac{3}{6}$$

Com o propósito de calcularmos $P\{A_2|A_1\}$ basta considerarmos apenas aquelas ocorrências experimentais para as quais ocorreu A_1. Como a primeira bola retirada não é devolvida à caixa, isto significa que consideramos apenas aqueles experimentos em que a primeira bola retirada foi uma das três bolas vermelhas. Assim, a segunda parte do experimento pode ser tratada como um novo experimento singular no qual uma bola deve ser retirada de uma caixa que contém duas bolas vermelhas, duas pretas e uma verde. Como resultado, $P\{A_2|A_1\}$ representa a probabilidade de se obter uma bola vermelha na retirada de uma bola da nova caixa de bolas. Aqui há cinco pontos do espaço amostral que representam a segunda retirada, como mostra a Figura 8: os dois primeiros

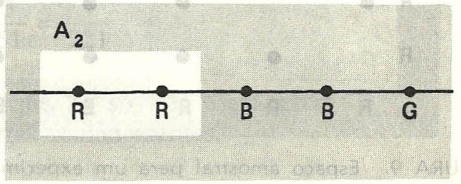


FIGURA 8. Espaço amostral da segunda retirada.

representando bolas vermelhas, os outros dois bolas pretas e o último uma bola verde. Donde, pela fórmula (2),

$$P\{A_2|A_1\} = \frac{2}{5}$$

A aplicação da fórmula (6) a estes dois resultados dá

$$P\{A_1 \text{ e } A_2\} = \frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5}$$

A vantagem de usarmos a fórmula (6) neste problema ficará evidente se tentarmos resolver este problema aplicando a fórmula (2) diretamente ao espaço amostral que corresponde a este experimento de dois estágios. O espaço amostral consistirá em trinta

pontos e assemelhar-se-á ao espaço amostral da Figura 6, onde a primeira retirada está simbolizada pelo retângulo inferior e a segunda, pelo retângulo da esquerda. Estarão faltando aqui os pontos da diagonal principal porque eles correspondem à obtenção do mesmo número de bolas coloridas em ambas as tentativas da retirada.

Este espaço bidimensional aparece na Figura 9. Usando a Figura 9 obtém-se a solução contando-se os pontos comuns a A_1 e A_2

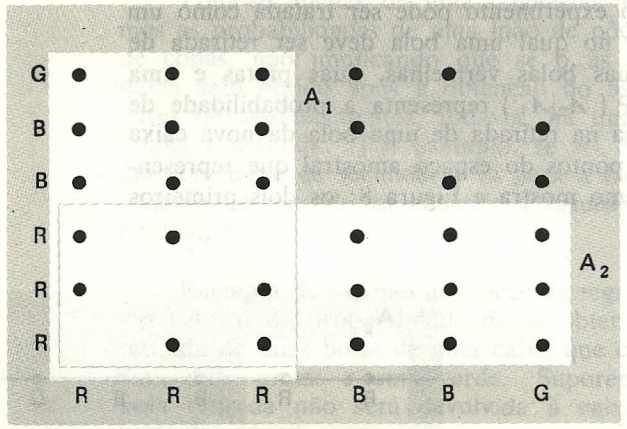


FIGURA 9. Espaço amostral para um experimento de dois estágios.

e dividindo-os pelo número total de pontos o que dá $6/30 = 1/5$. A vantagem da fórmula (6) é muito grande em experimentos de dois ou mais estágios. Daqui por diante, ao calcularmos probabilidades, as técnicas baseadas na fórmula (6) e nos experimentos de um único estágio serão usadas exclusivamente para evitarmos o gasto inútil de tempo devido ao fato de se aplicar a fórmula (2) diretamente ao espaço amostral do experimento todo. Deveríamos entender, contudo, que sempre será possível calcular qualquer tipo de probabilidade, trabalhando apenas com o espaço amostral original. O estudante deveria elaborar um problema deste tipo como teste de entendimento dos conceitos básicos.

Exemplo 2. Também como ilustração da regra de multiplicação, calcule a probabilidade de obter 2 prêmios ao abrir dois círculos, de uma mesa de círculos que contém 5 prêmios e 20 pontos em branco.

Se A_1 representar o evento da obtenção de um prêmio, na mesa de círculo, e A_2 o evento de ser premiado na segunda jogada, então a fórmula 6 dará

$$P\{A_1 \text{ e } A_2\} = \frac{5}{25} \cdot \frac{4}{24} = \frac{1}{30}$$

O valor de $P\{A_2|A_1\} = \frac{4}{24}$ surge do fato de que uma vez obtido o prêmio no primeiro círculo só restam 4 prêmios e 24 círculos não abertos

Exemplo 3. Como ilustração final, consideremos o problema de calcularmos a probabilidade de se obterem dois ases ao retirarmos duas cartas de um baralho de bridge.

Se A_1 representar o evento de se obter um ás numa primeira retirada e A_2 o evento de se obter um ás na segunda retirada, então a fórmula (6) dará

$$P\{A_1 \text{ e } A_2\} = \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}$$

O valor de $P\{A_2|A_1\} = 3/51$ porque há só três ases deixados após a primeira retirada das 51 cartas remanescentes.

7 — REGRA DA MULTIPLICAÇÃO PARA EVENTOS INDEPENDENTES

Se os eventos A_1 e A_2 são tais que a probabilidade da ocorrência de A_2 independe da ocorrência ou não ocorrência de A_1, então se diz que A_2 é independente de A_1 e pode-se escrever

$$P\{A_2|A_1\} = P\{A_2\}$$

Para este caso a regra de multiplicação reduz-se a

$$P\{A_1 \text{ e } A_2\} = P\{A_1\}P\{A_2\}$$

Visto que o evento (A_1 e A_2) é o mesmo que o evento (A_2 e A_1), A_1 e A_2 podem ser intercambiados na fórmula (6) para dar

$$P\{A_1 \text{ e } A_2\} = P\{A_2\}P\{A_1|A_2\}$$

A comparação dos segundos membros destas duas fórmulas mostra que $P\{A_1|A_2\} = P\{A_1\}$. Isto demonstra que A_1 é independente de A_2 quando A_2 é independente de A_1. Por causa desta independência mútua, é justo dizer que A_1 e A_2 são independentes, sem especificarmos qual dos dois independe do outro. Como resultado,

Quando A_1 e A_2 são independentes

$$P\{A_1 \text{ e } A_2\} = P\{A_1\}P\{A_2\} \tag{7}$$

Em vista deste resultado, pode-se afirmar que dois eventos são independentes se, e somente se, a probabilidade de sua ocorrência simultânea for igual ao produto de suas probabilidades individuais. Embora seja fácil estabelecer a condição que deve ser satisfeita se dois acontecimentos são independentes, não é sempre tão fácil na vida real decidir se dois acontecimentos são independentes.

Em jogos de azar, como roleta, sempre se supõe que jogadas consecutivas são acontecimentos independentes. Se alguém não quisesse aceitar tal suposição seria forçado a supor que a roleta é dotada de memória, ou que o operador da roda a estivesse manipulando secretamente.

Para determinar se dois acontecimentos da vida real A_1 e A_2 são independentes, é necessário fazer a seguinte pergunta: "Se A_1 ocorre, isto altera as chances de que A_2 ocorra se A_1 fosse completamente ignorado?" Se não se alteram as chances, os eventos são independentes.

A independência de eventos difere radicalmente do conceito de dois eventos mutuamente exclusivos. Nos problemas da vida real é muito fácil determinar se dois eventos A_1 e A_2 são mutuamente exclusivos. Tudo que se precisa fazer é perguntar: "Se A_1 ocorre, tal fato torna impossível a ocorrência de A_2?" Se a resposta é afirmativa, os dois eventos são mutuamente exclusivos.

Como ilustração de dois eventos mutuamente exclusivos, na vida real, façamos A_1 simbolizar o aumento do preço de uma obrigação municipal, no próximo mês, e A_2 o evento de que a taxa de rendimento desta obrigação também crescerá no próximo mês. Estes eventos são mutuamente exclusivos, porque, se o preço de uma obrigação aumentar, o rendimento da outra deverá decrescer e, portanto, se A_1 ocorrer, A_2 não poderá ocorrer. Como ilustração de dois eventos independentes, façamos A_1 ser o evento de que o preço

médio das obrigações municipais aumentarão, no próximo mês, e A_2 o evento segundo o qual o preço da prata vá aumentar, no próximo mês. Tais eventos podem ser tratados como independentes, porque não há evidência de que o preço da prata aumentará ou decrescerá com o aumento do preço médio das ações municipais.

Ao calcularmos probabilidades, é, muitas vezes, mais fácil calcularmos a probabilidade da não ocorrência de um evento, quando então subtraímos de 1 o resultado para obtermos o resultado de sua ocorrência. Por exemplo, para obtermos a probabilidade de conseguirmos pelo menos uma cara, no lançamento de três moedas, é mais fácil calcularmos a probabilidade de não obtermos cara, que é o mesmo que obtermos três coroas e, depois, subtrairmos o resultado de 1. Vemos pela Figura 1 que esta última probabilidade é 1/8; então, a probabilidade desejada é dada por $1 - \frac{1}{8} = \frac{7}{8}$.

8 — ILUSTRAÇÕES

Como ilustrações da aplicação das regras anteriores de probabilidade, consideremos alguns exemplos simples de cartas. Tiram-se duas cartas de um baralho comum de 52 cartas, sendo a primeira reposta antes da retirada da segunda.

Exemplo 1. Qual a probabilidade de que ambas as cartas venham a ser espadas? Seja A_1 o acontecimento "obtenção de espada na primeira extração" e A_2 o acontecimento "obtenção de espada na segunda". Visto que a primeira carta é reposta, a probabilidade de se obter uma espada na segunda extração não vai depender do fato de ter sido obtida ou não uma espada na primeira; desta forma pode-se supor A_2 independente de A_1. A fórmula (7) dará então:

$$P\{A_1 \text{ e } A_2\} = \frac{13}{52} \cdot \frac{13}{52} = \frac{1}{16}.$$

Exemplo 2. Qual é a probabilidade de as cartas serem ou duas espadas ou duas copas?

Seja B_1 o acontecimento "obtenção de 2 espadas" e B_2 o acontecimento "obtenção de 2 copas". Do resultado precedente segue-se que

$$P\{B_1\} = P\{B_2\} = \frac{1}{16}$$

Visto que os acontecimentos B_1 e B_2 são mutuamente *exclusivos* e o problema é calcular a probabilidade de que ou B_1 ou B_2 venha a ocorrer, aplica-se a fórmula (3). Assim:

$$P\{B_1 \text{ ou } B_2\} = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$$

Como anteriormente, façamos duas cartas serem extraídas de um baralho mas desta vez a primeira carta retirada não será reposta.

Exemplo 3. Qual a probabilidade de que ambas as cartas venham a ser espadas?

Aqui A_2 não é independente de A_1 porque se uma espada não é obtida na primeira extração, as chances de obtermos uma espada na segunda serão menores do que se uma não-espada tivesse sido obtida na primeira extração. Para este problema devemos usar a fórmula (6). Assim:

$$P\{A_1 \text{ e } A_2\} = \frac{13}{52} \cdot \frac{12}{51} = \frac{1}{17}$$

O segundo fator é 12/51 porque há apenas 51 cartas após a primeira extração, todas elas admitidas como possuindo a mesma chance de serem tiradas e ficarem somente 12 espadas.

Exemplo 4. Como ilustração que não envolve jogos de azar e que envolve mais de dois eventos independentes, consideremos o seguinte problema: supondo que a relação de crianças do sexo masculino seja 1/2 (o que só aproximadamente é verdade), ache a probabilidade de que numa família de 6 crianças (a) todas as crianças sejam do mesmo sexo, e (b) 5 crianças sejam homens e uma mulher.

(a) Seja A_1 o evento de que todas as crianças são homens e A_2 o evento de que todas as crianças são mulher. Porque A_1 e A_2 são eventos mutuamente exclusivos,

$$P\{A_1 \text{ ou } A_2\} = P\{A_1\} + P\{A_2\}$$

Como se pode supor que os seis nascimentos individuais sejam seis eventos independentes com relação ao sexo da criança, segue-se, usando a versão mais geral da fórmula (7) da multiplicação, que:

$$P\{A_1\} = P\{A_2\} = \left(\frac{1}{2}\right)^6$$

Donde

$$P\{A_1 \text{ ou } A_2\} = \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = \frac{1}{32}$$

(b) Seja A_1 o evento de que a criança mais velha seja menina e os outros meninos, A_2 o evento de que o segundo filho mais velho seja menina e os outros meninos e do mesmo modo para os eventos A_3, A_4, A_5, A_6. Como evento de se ter 5 homens e 1 mulher ocorrerá se. e somente se, ocorrer um dos seis eventos mutuamente exclusivos A_1, ..., A_6, segue-se da fórmula (3) que

$$P\{\text{5 homens e 1 mulher}\} = P\{A_1\} + \ldots\ldots + P\{A_6\}$$

Mas

$$P\{A_1\} = \ldots\ldots = P\{A_6\} = \left(\frac{1}{2}\right)^6$$

Donde

$$P\{\text{5 homens e 1 mulher}\} = 6\left(\frac{1}{2}\right)^6 = \frac{3}{32}$$

Exemplo 5. Há 25 pessoas num quarto. Qual a probabilidade de que, pelo menos, duas delas tenham a mesma data de aniversário? Suponha o ano de 365 dias.

Esta probabilidade pode ser obtida, primeiro, calculando a probabilidade de que 2 indivíduos não terão o mesmo aniversário e subtraindo depois o resultado de 1. O cálculo é feito como segue.

Escolha qualquer dos indivíduos e anote seu aniversário. Escolha um segundo indivíduo. A probabilidade de que seu aniversário difira do aniversário dos dois primeiros indivíduos selecionados é de 363/365. Continua este processo até que todos os 25 indivíduos sejam escolhidos. Supõe-se que estes eventos individuais sejam independentes; portanto, a probabilidade de que 2 indivíduos não tenham o mesmo aniversário é dada por

$$\frac{364}{365} \cdot \frac{363}{365} \cdot \ldots\ldots \frac{341}{365} = 0{,}43$$

Donde a probabilidade desejada é dada por

$$1 - 0{,}43 = 0{,}57$$

Tal resultado é surpreendentemente grande, e indivíduos que não viram cálculos de probabilidade não acreditarão no resultado.

Portanto pode-se fazer uma boa aposta entre amigos, se houver um número razoável deles. Se o número de indivíduos for 30, esta probabilidade aumenta para 0,71.

Exemplo 6. Suponhamos que um míssil dirigível tenha 5 seções de circuitos distintos através dos quais deve passar um sinal para que ele funcione com exatidão. Cada uma das seções individuais tem 2 circuitos através dos quais deve passar um sinal, pelo menos um desses deve funcionar se o sinal tiver de atravessar a seção. Se a probabilidade de que qualquer circuito singular falhe é de 0,2, qual a probabilidade de que um sinal passe (a) através da primeira seção, e (b) através de todas as 5 seções?

(a) Se A_6 é o evento de que o primeiro circuito, numa dada seção, opera adequadamente e A_2 o evento de que o segundo circuito opera adequadamente, então pode-se supor que A_1 e A_2 são eventos independentes. A probabilidade requerida aqui é $P\{A_1 \text{ ou } A_2\}$, isto é, a probabilidade de que pelo menos um dos circuitos opera adequadamente. Isto é obtido mais facilmente calculando-se a probabilidade de que nem A_1 nem A_2 ocorra, que é $(0{,}2)(0{,}2) = 0{,}4$. O complemento deste evento é

$$P\{A_1 \text{ ou } A_2\} = 1 - 0{,}4 = 0{,}96$$

(b) como cada seção de circuito opera independentemente das outras a probabilidade de todas as cinco seções operarem adequadamente é dada pelo produto das probabilidades, que é

$$(0{,}96)^5 = 0{,}815$$

Embora as regras anteriores de probabilidade sejam derivadas do pressuposto de que se espera que todas as ocorrências possíveis do experimento em questão ocorram com relativa freqüência, as regras são válidas para experimentos mais gerais. Podem mesmo ser aplicadas a eventos relacionados a experimentos onde entra um número infinito de ocorrências possíveis. Estes experimentos mais gerais serão considerados nos Capítulos 4 e 5.

9 — FÓRMULA DE BAYES

Há uma certa classe de problemas baseados na aplicação da fórmula (5) que leva a cálculos complexos; portanto é conveniente ter uma fórmula para se resolver tais problemas de maneira sistemática. Tais problemas podem ser ilustrados por este tipo de problema acadêmico. Suponhamos uma caixa que contém 2 bolas vermelhas e uma bola branca, e uma segunda caixa que contém 2 bolas vermelhas e 2 bolas brancas. Escolhe-se uma das caixas ao acaso e retira-se dela uma bola. Se a bola retirada for vermelha, qual a probabilidade de que ela tenha vindo da primeira caixa? Façamos A_1 significar o evento "escolha da primeira caixa" e A_2 o evento "retirada da bola vermelha". Então o problema é calcular a probabilidade condicional $P\{A_1|A_2\}$. Isto será feito pela fórmula (5) com A_1 e A_2 intercambiados nesta fórmula. Como a expressão por acaso significa que cada caixa tem a mesma probabilidade de ser escolhida, segue-se que a probabilidade da retirada da primeira caixa é 1/2, o mesmo acontecendo com a probabilidade da retirada da segunda caixa. O cálculo do numerador na fórmula (5) pode ser realizado usando-se a fórmula (6) na ordem dos eventos, de acordo com a marcação da tabela. Assim:

$$P\{A_2 \text{ e } A_1\} = P\{A_1 \text{ e } A_2\} = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$$

O denominador, $P\{A_2\}$, pode ser calculado considerando-se as duas maneiras mutuamente exclusivas em que A_2 pode ocorrer, isto é, obtendo-se a primeira caixa e depois a bola vermelha ou obtendo-se a segunda caixa e depois a bola vermelha. Pela fórmula (3), $P\{A_2\}$ será dado pela soma das probabilidades destas duas probabilidades mutuamente exclusivas; donde

$$P\{A_2\} = \frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{2}{4} = \frac{7}{12}$$

A aplicação da versão modificada da fórmula (5) fornece então o resultado desejado, isto é:

$$P\{A_1|A_2\} = \frac{1}{3} \Big/ \frac{7}{12} = \frac{4}{7}$$

Este problema poderia ser resolvido muito facilmente considerando-se o espaço amostral do experimento; todavia, o objetivo aqui é trabalhar com a fórmula (5) e tentar a obtenção de uma fórmula para o tratamento de problemas mais complicados do tipo que estamos vendo.

O problema precedente é um caso especial de problemas do seguinte tipo. Apresenta-se um experimento de duas etapas. A primeira etapa pode ser descrita afirmando-se que exatamente uma das k ocorrências possíveis deve ocorrer quando se realizar o experi-

mento completo. Estas ocorrências possíveis serão simbolizadas por $e_1, e_2, \ldots, e_k$. Na segunda etapa há, digamos, m ocorrências possíveis, exatamente uma das quais deve ocorrer. Serão simbolizadas por $o_1, o_2, \ldots, o_m$. Os valores das probabilidades para cada uma das ocorrências possíveis $e_1, e_2, \ldots, e_k$ são dados. Como antes, as ocorrências serão simbolizadas por $P\{e_1\}, P\}e_2, \ldots, P\{e_k\}$. Os valores de todas as probabilidades condicionais do tipo $P\{o_j|e_i\}$ que representa a probabilidade da ocorrência do evento o_j da segunda etapa quando se sabe que ocorreu o evento e_i da primeira etapa, são também dados.

O problema é calcular a probabilidade de que o evento e_i da primeira etapa ocorreu quando se sabe que ocorreu o evento o_j da segunda etapa. Esta probabilidade condicional seria escrita $P\{e_1|o_1\}$; os cálculos para qualquer outro par seriam os mesmos.

Em termos da notação atual a fórmula (5) assume a seguinte feição:

$$P\{e_1|o_1\} = \frac{P\{e_1 \text{ e } o_1\}}{P\{o_1\}} \tag{8}$$

A fórmula (6) dá

$$P\{e_1 \text{ e } o_1\} = P\{e_1\}P\{o_1|e_1\} \tag{9}$$

Das informações dadas neste problema observa-se que as duas probabilidades do segundo membro da equação (9) são conhecidas; portanto, o numerador da fórmula (8) pode ser obtido da fórmula (9). O valor de $P\{o_1\}$ na fórmula (8) pode ser calculado considerando-se todas as maneiras mutuamente exclusivas pelas quais o_1 pode ocorrer em conjunto com a primeira etapa do experimento. O evento o_1 da segunda etapa ocorrerá se ocorrer o evento e_1 da primeira etapa e também o_1, ou se o evento e_2 da primeira etapa ocorrer e também novamente $o_1, \ldots$ ou se o evento e_k da primeira etapa ocorrer e também, finalmente, o_1.

Donde:

$$P\{o_1\} = P\{e_1 \text{ e } o_1\} + P\{e_2 \text{ e } o_1\} + \cdots + P\{e_k \text{ e } o_1\}$$

Se e_1 for substituído por e com a subscrição apropriada na fórmula (9), a fórmula poderá ser usada para se calcular a probabilidade de cada uma dessas possibilidades mutuamente exclusivas.

A aplicação da fórmula (3) a esses k eventos mutuamente exclusivos fornece a fórmula

$$P\{o_1\} = P\{e_1\}P\{o_1|e_1\} + P\{e_2\}P\{o_1|e_2\} + \cdots + P\{e_k\}P\{o_1|e_k\}$$

que pode ser escrita, de forma condensada, usando-se o símbolo Σ de somatória, como segue:

$$P\{o_1\} = \sum_{i=1}^{k} P\{e_i\}P\{o_1|e_i\}$$

Aqui e_i significa o evento número i no conjunto $e_1, e_2, \ldots, e_k$ e o símbolo de somatória nos diz para somarmos todas as probabilidades, começando com e_1 e terminando com e_k. Este resultado juntamente com a fórmula (9) quando aplicado à fórmula (8) dará a fórmula desejada, conhecida por fórmula de Bayes

(10) Fórmula de Bayes.

$$P\{e_1|o_1\} = \frac{P\{e_1\}P\{o_1|e_1\}}{\sum_{i=1}^{k} P\{e_i\}P\{o_1|e_i\}}$$

Voltando ao problema que foi resolvido antes sem esta fórmula, pode-se observar que houve dois eventos e_1 e e_2 na primeira etapa, o que corresponde à escolha da primeira ou da segunda caixa e que $P\{e_1\} = P\{e_2\} = \frac{1}{2}$. A segunda etapa também consistia em dois eventos o_1 e o_2, o que corresponde à obtenção de uma bola vermelha ou branca. As probabilidades condicionais de se obter uma bola vermelha, baseadas no que transpirou da primeira etapa, foram dadas por

$$P\{o_1|e_1\} = \tfrac{2}{3} \text{ e } P\{o_1|e_2\} = \tfrac{2}{4}$$

Será possível observar que a substituição destes valores na fórmula (10) fornece o resultado obtido antes.

Exemplo. Consideremos agora uma aplicação mais prática desta fórmula. Suponhamos que se realizou um teste para se detectar determinada doença rara e que este teste seja capaz de descobrir a doença em 97 por cento das pessoas afetadas. Suponhamos que,

além disso, quando ele é experimentado em pessoas sadias, 5 por cento delas são incorretamente diagnosticadas como tendo a doença. Finalmente, suponhamos que, quando é tentado o teste em indivíduos que tenham outras doenças mais brandas, 10 por cento deles são incorretamente diagnosticados.

Sabe-se que as porcentagens, na população total, dos indivíduos dos três tipos considerados aqui, são de 1 por cento, 96 por cento e 3 por cento, respectivamente. O problema é calcular a probabilidade de que um indivíduo, escolhido ao acaso, da população, e testado com relação à doença rara, tenha realmente a doença, se o teste indicar que ele está afetado.

Há aqui três eventos e_1, e_2 e e_3 na primeira etapa, correspondentes aos três tipos de indivíduos da população. Suas probabilidades correspondentes são $P\{e_1\} = 0{,}01$, $P\{e_2\} = 0{,}96$ e $P\{e_3\} = 0{,}03$. Há dois eventos o_1 e o_2 na segunda etapa correspondendo a saber se o objetivo do teste é mostrar que o indivíduo tem a doença. As probabilidades condicionais são dadas por $P\{o_1|e_1\} = 0{,}97$, $P\{o_1|e_2\} = 0{,}05$ e $P\{o_1|e_3\} = 0{,}10$. Em termos da notação presente, o problema é calcular $P\{e_1|o_1\}$. Uma aplicação direta da fórmula (10) baseada nas probabilidades precedentes dará a resposta, a saber:

$$P\{e_1|o_1\} = \frac{(0{,}01)\,(0{,}97)}{(0{,}01)\,(0{,}97)+(0{,}96)\,(0{,}5)+(0{,}3)\,(0{,}10)} = 0{,}16$$

Este resultado pode parecer um tanto surpreendente porque mostra que somente 16 por cento dos indivíduos indicados pelo teste como possuidores da doença realmente possuirão a doença, se o teste for aplicado à população toda. Os 84 por cento que foram falsamente diagnosticados sentirão a angústia mental temporária causada pela crença de que eles têm a doença, antes que testes ulteriores revelem a falsidade do diagnóstico. Sentirão também o fato de terem que se submeter a outros testes quando aparecer claro que aqueles testes não foram necessários. Um cálculo como o precedente devia fazer as autoridades ponderarem mais, antes de defenderem testes de massa.

10 — TÉCNICAS DE CONTAGEM — ÁRVORES

Na solução de alguns problemas difíceis onde entram dois ou mais estágios, é conveniente termos métodos sistemáticos para o cálculo de eventos compostos. Um método gráfico que se mostrou

muito útil, no caso, é o que se baseia na chamada *árvore de probabilidade*.

Com o intuito de mostrarmos como se constrói esta árvore, consideremos, mais uma vez, o primeiro exemplo ao qual a fórmula (6) foi aplicada. Uma caixa contém três bolas vermelhas, duas pretas e uma verde, e duas destas bolas devem ser retiradas. Este é um experimento de duas etapas para o qual as várias possibilidades que podem ocorrer podem ser representadas por uma árvore horizontal como a que aparece na Figura 10. Cada etapa de um experimento de etapas múltiplas possui tantos ramos quantas forem as possibilidades, neste estágio.

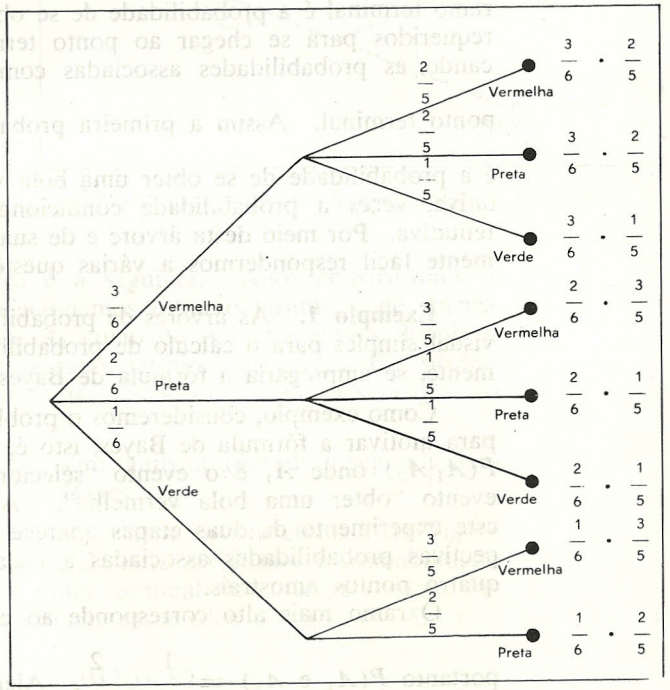


FIGURA 10. Árvore de probabilidade.

Aqui, há três ramos principais na primeira etapa e três ramos secundários para cada ramo, na segunda etapa, exceto para o último ramo, onde só há dois ramos, porque é impossível obter-se uma bola verde, na segunda retirada, se se obtém essa bola verde na primeira retirada. O número total de ramos terminais desta árvore dá o

número total de ocorrências possíveis do experimento composto e, portanto, os pontos terminais desses ramos podem ser tratados como pontos amostrais de um espaço amostral.

A probabilidade associada a cada ramo da árvore é a probabilidade condicional de que o evento sob este ramo ocorrerá, sujeita (a probabilidade) à condição de que os eventos dos ramos precedentes ocorrerão todos. Assim, os 2/5 colocados sobre o topo do ramo terminal são a probabilidade condicional de se obter uma bola vermelha, na segunda retirada se se obtiver uma bola vermelha na primeira tentativa. A probabilidade colocada no término de um ramo terminal é a probabilidade de se obter a seqüência de eventos requeridos para se chegar ao ponto terminal e obtém-se multiplicando as probabilidades associadas com os ramos que levam ao ponto terminal. Assim a primeira probabilidade terminal $\frac{3}{6} \cdot \frac{2}{5}$ é a probabilidade de se obter uma bola vermelha, na primeira tentativa, *vezes* a probabilidade condicional de obtê-la, na segunda tentativa. Por meio desta árvore e de suas probabilidades é relativamente fácil respondermos a várias questões sobre probabilidade.

Exemplo 1. As árvores de probabilidade fornecem um método visual simples para o cálculo de probabilidades para as quais, geralmente, se empregaria a fórmula de Bayes.

Como exemplo, consideremos o problema anterior que foi usado para motivar a fórmula de Bayes, isto é, o problema de se calcular $P(A_1|A_2)$ onde A_1 é o evento "selecionar a caixa 1" e A_2 é o evento "obter uma bola vermelha". A árvore correspondente a este experimento de duas etapas aparece na Figura 11 com as respectivas probabilidades associadas a cada ramo e a cada um dos quatro pontos amostrais.

O ramo mais alto corresponde ao evento composto A_1 e A_2, portanto $P(A_1 \text{ e } A_2) = \frac{1}{2} \cdot \frac{2}{3}$. Além disso, segue-se da definição (1) que $P(A_2) = \frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{2}{4}$, porque o evento A_2 consiste em dois pontos amostrais associados à palavra "Vermelho". Donde, pela fórmula (5),

$$P\{A_1|A_2\} = \frac{P\{A_1 \text{ e } A_2\}}{P\{A_2\}} = \frac{\frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{2}{4}} = \frac{4}{7}$$

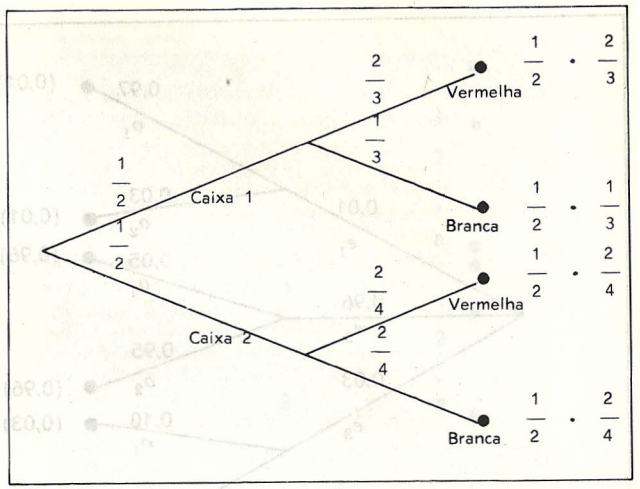


FIGURA 11. Árvore de um problema bayesiano.

Vê-se, agora, que a técnica é a seguinte. Após construirmos a árvore de probabilidade, selecionamos o ramo terminal que corresponde à ocorrência tanto de A_1 como de A_2. Dividimos depois a probabilidade associada a este ramo terminal pela soma das probabilidades de todos os ramos terminais que estão associados ao evento A_2.

Exemplo 2. Esta técnica será também aplicada à segunda ilustração associada à fórmula de Bayes.

Primeiro, é necessário construir uma árvore como a da Figura 12. Para se obter $P\{e_1|o_1\}$ basta agora dividirmos a probabilidade (0,01) (0,97) associada ao ramo terminal de cima pela soma das probabilidades dos ramos terminais associados à letra o_1. Assim

$$P\{e_1|o_1\} = \frac{(0{,}01)\ (0{,}97)}{(0{,}01)\ (0{,}97)+(0{,}96)\ (0{,}05)+(0{,}03)\ (0{,}10)} = 0{,}16$$

11 — TÉCNICAS DE CONTAGEM — COMBINAÇÕES

Se houver muitas etapas num experimento e várias possibilidades em cada etapa, a árvore de probabilidades associada ao experimento tornar-se-ia grande demais para ser manipulada. Para tais problemas a contagem de pontos amostrais é simplificada por meio

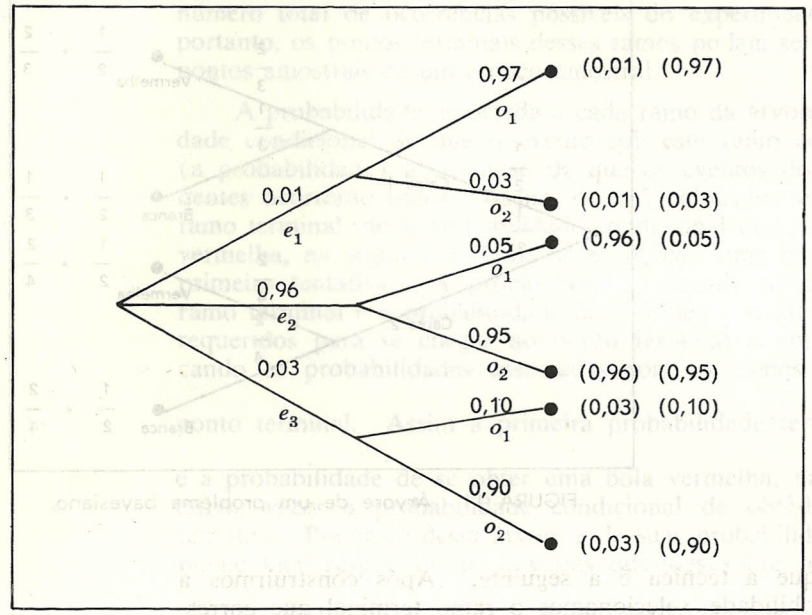


FIGURA 12. Árvore de um problema bayesiano.

de fórmulas algébricas. Com este objetivo, consideremos um experimento de duas etapas para o qual há r possibilidades na primeira etapa e s possibilidades na segunda. Uma árvore representando este experimento aparece na Figura 13. Como cada um dos ramos principais de r tem ramos terminais s ligados a si, o número total de possibilidades aqui é rs, e portanto este seria o número de pontos amostrais, no espaço amostral, para este experimento de duas etapas. Se acrescentarmos uma terceira etapa com t possibilidades, o número total de pontos amostrais tornar-se-ia rst. É evidente que isto pode ser estendido para qualquer número de etapas.

Agora, consideremos a aplicação deste método de contagem ao problema de determinarmos de quantas maneiras é possível selecionarmos r objetos de n objetos distintos. Para este fim, consideremos primeiro o problema particular de determinarmos quantas palavras de três letras podem ser formadas com as cinco letras a, b, c, d, e, se só pudermos usar cada letra uma vez, em cada palavra e se cada conjunto de três letras for chamado uma palavra de três letras.

A formação de uma palavra de três letras pode ser considerada como um experimento de três etapas no qual a primeira etapa

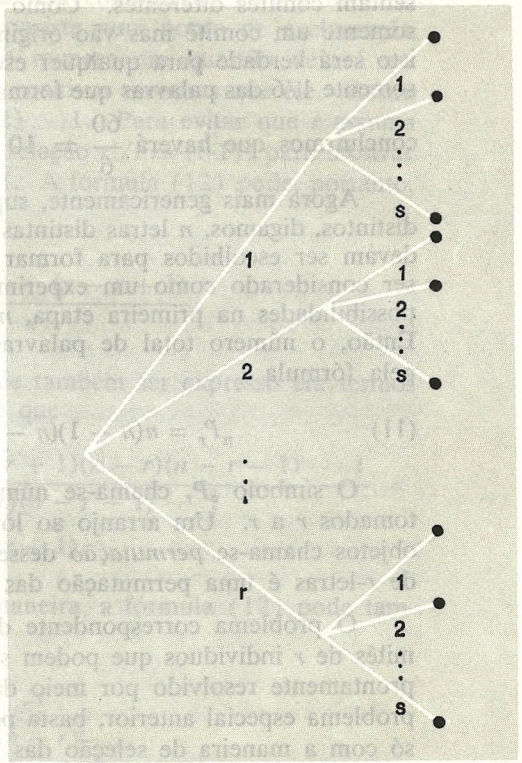


FIGURA 13. Árvore para um experimento de dois estágios.

consiste na escolha da primeira letra da palavra. Neste experimento há cinco possibilidades para a primeira etapa, mas só quatro para a segunda, porque a letra escolhida na primeira etapa não está mais à disposição na etapa seguinte. Há somente três possibilidades na terceira etapa; portanto, conforme o método de contagem precedente, o número total de palavras de três letras que pode ser formado é dado por $5.4.3 = 60$.

O problema anterior poderá ser um pouco modificado pela pergunta: "Quantos comitês de três indivíduos podem ser escolhidos de um grupo de cinco indivíduos?" Se as letras a, b, c, d, e forem associadas aos cinco indivíduos, então a palavra de três letras corresponderá a um comitê de três; todavia, duas palavras que empreguem as mesmas três letras, em ordem diversa, corresponderá ao mesmo comitê. Por exemplo, *cis* e *sic* são palavras distintas mas não repre-

sentam comitês diferentes. Como as três letras c, i e s produzirão somente um comitê mas vão originar $3.2.1 = 6$ letras diversas, e isto será verdade para qualquer escolha de três letras, segue-se que somente $1/6$ das palavras que formam letras formarão comitês; donde concluirmos que haverá $\frac{60}{6} = 10$ de tais comitês.

Agora mais genericamente, suponhamos que se tenha n objetos distintos, digamos, n letras distintas $a, b, c \ldots$ e que r desses objetos devam ser escolhidos para formar palavras de r-letras. Isto pode ser considerado como um experimento de r etapas, no qual há n possibilidades na primeira etapa, n-1 possibilidades na segunda etc. Então, o número total de palavras que pode ser formado é dado pela fórmula

$$(11) \qquad {}_nP_r = n(n-1)(n-2)\cdots(n-r+1)$$

O símbolo ${}_nP_r$ chama-se número de permutação de n objetos tomados r a r. Um arranjo ao longo da linha de um conjunto de objetos chama-se *permutação* desses objetos; portanto, uma palavra de r-letras é uma permutação das r letras usadas para construí-la.

O problema correspondente da contagem dos números de comitês de r indivíduos que podem ser selecionados de n indivíduos é prontamente resolvido por meio da fórmula precedente. Como no problema especial anterior, basta perceber que o comitê se relaciona só com a maneira de seleção das r letras e não com a maneira de seu arranjo. Com r letras distintas podem ser arranjadas, ao longo de uma linha, para produzir $r(r\text{-}1)\ (r\text{-}2)\ldots 1$ palavras distintas, mas só um comitê, o número de palavras deve ser dividido por $r(r\text{-}1)\ (r\text{-}2)\ldots 1$ para dar o número de comitês distintos. Assim, o número de comitês distintos que podem ser realizados é dado pela fórmula

$$(12) \qquad \binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r(r-1)(r-2)\cdots 1}$$

O símbolo $\binom{n}{r}$ é chamado número de combinações de n objetos tomados r a r. Uma *combinação* de r objetos é uma simples relação de r objetos sem consideração pela ordem em que são selecionados ou arrumados após a seleção. Uma permutação de r objetos pode ser considerada como a ocorrência de um experimento de duas etapas no qual a primeira etapa consiste na escolha da combinação de r objetos e no arranjo desta combinação ao longo de uma linha.

Um símbolo útil para se usar em combinação com a fórmula (12) é o símbolo *fatorial*, que consiste num ponto de exclamação após o número, indicando que o número em questão deverá ser multiplicado por todos os inteiros positivos menores que ele. Assim $4! = 4.3.2.1$ e $r! = r(r\text{-}1)\ (r\text{-}2)\ldots 1$. Para evitar que r assuma o valor zero, basta fazer $r = 1$ na relação $r!/r = (r\text{-}1)!$ para motivar a definição de que 0! seja igual a 1. A fórmula (12) pode, portanto, ser escrita da seguinte maneira

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}$$

O numerador desta expressão pode também ser expresso em termos de notação fatorial observando-se que

$$\frac{n!}{(n-r)!} = \frac{n(n-1)\cdots(n-r+1)(n-r)(n-r-1)\cdots 1}{(n-r)(n-r-1)\cdots 1}$$
$$= n(n-1)\cdots(n-r+1)$$

Usando símbolos fatoriais desta maneira, a fórmula (12) pode também ser escrita compactamente:

$$(13) \qquad \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Exemplo 1. A utilidade da fórmula (12) ou (13), para fins de contagem, será ilustrada no cálculo da probabilidade de se obter cinco espadas em uma mão de cinco cartas tiradas de um baralho de 52 cartas.

Aqui, o número total de ocorrências possíveis corresponde ao número total de mãos de cinco cartas que podem ser formadas de um baralho de 52 cartas distintas. Como o arranjo não tem interesse aqui, trata-se de um problema de combinação. O número total de mãos possíveis é, conforme a fórmula (12), dado por

$$\binom{52}{5} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

O número de sucessos que correspondem à ocorrência do evento desejado é igual ao número de maneiras de selecionarmos cinco espadas de treze espadas. É dado por

$$\binom{13}{5} = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

A probabilidade desejada é dada pela razão destes dois números. É igual a

$$\frac{13.12.11.10.9}{52.51.50.49.48} = \frac{33}{66640} = 0{,}0005$$

Moral da história: parece que não se deve esperar obter uma mão de pôquer de cinco cartas que só contenha espadas. Mesmo se obtivermos uma mão que contém cinco cartas do mesmo naipe a probabilidade é apenas quatro vezes maior, isto é, 0,002, o que é ainda desencorajadoramente pequena.

Exemplo 2. Como outra ilustração do uso destas fórmulas, consideremos mais este problema. Se uma lista de 20 indivíduos que se ofereceram como voluntários para doar sangue, quando necessário para transfusão, possuir 15 indivíduos com sangue tipo **B** e se 3 indivíduos forem escolhidos, ao acaso, desta lista, qual a probabilidade de (a) todos os três serem do tipo **B**; (b) dois serem do tipo **B** e um não; (c) pelo menos um ser do tipo **B**?

(a) O número total de ocorrências possíveis é o número total de maneiras de se escolher 3 indivíduos de um grupo de 20 indivíduos dado por $\binom{20}{3}$. O número de sucessos que corresponde à ocorrência do evento desejado é o número de maneiras de se escolher 3 indivíduos de um grupo de 15 indivíduos dado por $\binom{15}{3}$. A probabilidade de ocorrência deste evento favorável é, portanto, dada pela razão

$$\frac{\binom{15}{3}}{\binom{20}{3}} = \frac{15.14.13}{20.19.18} = \frac{91}{228} = 0{,}40$$

(b) O número de maneiras de se escolher 2 indivíduos de um grupo de tipo **B** e 1 indivíduo de um ou outro grupo pode ser calculado da seguinte maneira. Como 2 indivíduos devem ser selecionados de um grupo de 15 indivíduos, há $\binom{15}{2}$ maneiras de se fazer esta seleção. Do mesmo modo, como 1 indivíduo deve ser selecionado de um grupo de 5 indivíduos, há $\binom{5}{1}$ maneiras de se fazer esta seleção. Mas pelo princípio fundamental de contagem explicado no primeiro parágrafo desta seção e mostrado na Figura 13, o número total de modos de se realizar estas duas seleções é dado pela multiplicação do número de modos de se realizar as escolhas separadas deste experimento de duas etapas. Portanto, o número total de maneiras de se fazer as seleções desejadas é dado por

$$\binom{15}{2}\binom{5}{1}$$

Este é o número total dos pontos amostrais que corresponde à ocorrência do evento desejado; donde, a probabilidade de que exatamente 2 do grupo possuam sangue do tipo **B** e 1 não o possua, é dada pela razão

$$\frac{\binom{15}{2}\binom{5}{1}}{\binom{20}{3}} = \frac{15.14.5.3}{20.19.18} = \frac{35}{76} = 0{,}46$$

(c) Este evento ocorrerá se não ocorrer o evento "obter 0 (zero) destes indivíduos". A última probabilidade é dada por $\binom{5}{3}/\binom{20}{3}$. A probabilidade desejada é, portanto, dada por

$$1 - \frac{\binom{5}{3}}{\binom{20}{3}} = 1 - \frac{5.4.3}{20.19.18} = 1 - \frac{1}{114} = \frac{113}{114} = 0{,}99$$

12 — ILUSTRAÇÕES ADICIONAIS

Como aconteceu com relação aos capítulos precedentes e acontecerá com relação aos seguintes, esta última seção destina-se a uma revisão das técnicas de solução dos problemas apresentados no capítulo. Para servir como revisão, o estudante deverá tentar resolver estes problemas antes de olhar as soluções. Como servem de ilustrações adicionais à medida que avançamos de uma para outra seção, o número da seção associado a cada parte do problema indicará a que seção pertence o problema.

Exemplo 1. Dois dados foram alterados colocando-se 2 pontos na face do dado que deveria possuir um ponto apenas. Como resultado cada dado conterá duas faces com 2 pontos e nenhuma com 1 ponto. Os 2 dados são então lançados, ao mesmo tempo. Supondo-se que se possa distinguir os dois dados, resolva os seguintes problemas: (a) 2 — Construa o espaço amostral do experimento. (b) 3 — Atribua probabilidades aos pontos do espaço amostral. (c) 4 — Usando a definição (1) calcule a probabilidade (i) de se obter um total de 6 pontos nos dois dados, e (ii) que pelo menos um dos dados mostre um 2. Seguem-se as soluções.

(a) O espaço amostral está convenientemente representado pelos seguintes 25 pontos que foram marcados para indicar a ocorrência, em cada dado.

.	.	.	.	
22	32	42	52	62
.	.	.	.	.
23	33	43	53	63
.	.	.	.	.
24	34	44	54	64
.	.	.	.	.
25	35	45	55	65
.	.	.	.	.
26	36	46	56	66

(b) Supondo-se que os dois dados se comportem da mesma maneira que dois dados normais para os quais o espaço amostral é dado no Quadro 1 e para os quais a cada ponto foi atribuída a probabilidade de 1/36, as probabilidades a serem atribuídas aqui deveriam estar em concordância com aquelas probabilidades. Donde, como o evento 22 aqui corresponde ao evento composto consistente dos eventos simples 11, 12, 21 e 22 para o Quadro 1, deve-se atribuir a probabilidade 4/36 ao ponto marcado com 22, no atual espaço amostral. Cada um dos pontos restantes deste espaço que tenha um 2 em sua marca deve receber a probabilidade 2/36, porque há dois eventos simples da tabela 1 que os produz. A todos os demais pontos deve-se atribuir a probabilidade 1/36, porque não diferem dos correspondentes do Quadro 1.

(c) (i) Se A simbolizar este evento vê-se examinando o espaço amostral em (a) que os três pontos 24, 33 e 42 constituem o evento composto A; donde, aplicando a definição (1) e usando as probabilidades atribuídas a estes pontos em (b), segue-se que

$$P\{A\} = \frac{2}{36} + \frac{1}{36} + \frac{2}{36} = \frac{5}{36}.$$

(ii) Se B simbolizar este evento, vê-se que B consiste nos eventos simples 22, 23, 24, 25, 26, 32, 42, 52 e 62. De (b) o ponto 22 recebeu a probabilidade 4/36 e os pontos restantes a probabilidade 2/36; conseqüentemente, como há oito dos últimos,

$$P\{B\} = \frac{4}{36} + 8\left(\frac{2}{36}\right) = \frac{5}{9}$$

Exemplo 2. Uma bolsa contém 2 moedas de um centavo (p), três de 10 centavos (n) e 4 de um cruzeiro (d). Duas moedas serão escolhidas da bolsa, ao acaso. Usando as fórmulas de adição e multiplicação, calcule a probabilidade de que

(a) 6. ambas as moedas sejam de 1 centavo;

(b) 6. uma moeda seja de 1 centavo e as outras de 1 cruzeiro;

(c) 6. ambas as moedas sejam da mesma espécie;

(d) 6. nenhuma moeda seja de 10 centavos;

(e) 6. pelo menos se consiga uma moeda de 10 centavos.

Eis as soluções:

(a) aplicando-se a fórmula (6) e considerando o experimento em duas etapas temos

$$P\{\text{pp}\} = \frac{2}{9} \cdot \frac{1}{8} = \frac{1}{36}.$$

(b) Este evento ocorrerá se um ou outro dos dois eventos mutuamente exclusivos ocorrer pd ou dp. A aplicação das fórmulas (6) e (3) dá então, o resultado

$$P\{\text{pd}\} + P\{\text{dp}\} = \frac{2}{9} \cdot \frac{4}{8} + \frac{4}{9} \cdot \frac{2}{8} = \frac{2}{9}$$

(c) Este evento ocorrerá se um dos três eventos mutuamente exclusivos pp, nn, ou dd ocorrer. Aqui a resposta é dada por

$$P\{\text{pp}\} + P\{\text{nn}\} + P\{\text{dd}\} = \frac{2}{9} \cdot \frac{1}{8} + \frac{3}{9} \cdot \frac{2}{8} + \frac{4}{9} \cdot \frac{3}{8} = \frac{5}{18}$$

(d) Há 3 moedas de 10 centavos e 6 que não são de dez centavos. A probabilidade de se obter 2 moedas que não sejam de 10 centavos é $\frac{6}{9} \cdot \frac{5}{8} = \frac{5}{12}$

(e) A probabilidade de se obter, pelo menos, uma moeda de 10 centavos é igual a 1 menos a probabilidade de se obter 2 moedas que não sejam de 10 centavos. Considerando (d) a resposta deve ser $1 - \frac{5}{12} = \frac{7}{12}$

Exemplo 3. 6 — Suponhamos que 80 por cento dos compradores de carros usados sejam bons pagadores.

Suponhamos, além disso, que haja uma probabilidade de 0,7 de que um bom pagador obtenha cartão de crédito e que esta probabilidade passe a ser de apenas 0,4 por cento para um mau pagador.

Calcule a probabilidade de que

(a) um comprador de carro selecionado ao acaso tenha um cartão de crédito;

(b) um comprador de carro selecionado ao acaso e que tenha um cartão de crédito seja um bom pagador;

(c) um comprador de carro escolhido ao acaso e que não tenha cartão de crédito seja um bom pagador.

Seguem-se as soluções.

(a) Façamos A significar o evento "seleção de um bom pagador" e B o evento "seleção de um indivíduo que tenha um cartão de crédito". A probabilidade desejada é dada então por $P\{B\}$. Se $\bar{A}$ significar o evento "A não ocorrerá", então B ocorrerá se A ocorrer e depois B ocorrer, ou se $\bar{A}$ ocorrer e depois B ocorrer. Donde

$$P\{B\} = P\{A\}\ P\{B|A\} + P\{\bar{A}\}\ P\{B|\bar{A}\}$$
$$= (0,8)\ (0,7) + (0,2)\ (0,4) = 0,64$$

(b) Em termos da notação em (a), a probabilidade desejada aqui é dada por $P\{A|B\}$. Usando o resultado em (a)

$$P\{A|B\} = \frac{P\{A \text{ e } B\}}{P\{B\}} = \frac{P\{A\}\ P\{B|A\}}{P\{B\}}$$

$$= \frac{(0,8)\ (0,7)}{0,64} = \frac{7}{8}$$

(c) A probabilidade desejada aqui é dada por $P\{A|\bar{B}\}$. Usando o resultado em (a), mais uma vez

$$P\{A|\bar{B}\} = \frac{P\{A \text{ e } \bar{B}\}}{P\{\bar{B}\}} = \frac{P\{A\}\ P\{\bar{B}|A\}}{1 - P\{B\}}$$

$$= \frac{(0,8)\ (0,3)}{0,36} = \frac{2}{3}$$

Exemplo 4. 9 — Um aparelho para testes de válvulas de rádio pode sempre detectar uma válvula ruim mas em 2 por cento das vezes que ele indica válvula estragada, de fato, a válvula está boa. Se 97 por cento das válvulas são boas, qual a probabilidade de que, se uma nova válvula for selecionada, ao acaso, e o aparelho de teste afirmar que é uma válvula estragada, esta válvula seja, de fato, boa?

Seja e_1 o evento "obter uma boa válvula" e e_2 o evento "obter uma válvula ruim". Então $P\{e_1\} = 0,97$ e $P\{e_2\} = 0,03$. Seja o_1 o evento "o aparelho afirma que a válvula testada é ruim" e o_2 o evento "a válvula testada é boa". Então $P\{o_1|e_1\} = 0,02$ e $P\{o_1|e_2\} = 1$.

A solução do problema é dada pelo valor de $P\{e_1|o_1\}$. Usando a fórmula (10) e os valores precedentes teremos:

$$P\{e_1|o_1\} = \frac{(0,97)\ (0,02)}{(0,97)\ (0,2) + (0,03)\ 1} = 0,39$$

Donde, cerca de 39 por cento das vezes que uma nova válvula for declarada má pelo aparelho, ela será realmente boa.

Exemplo 5. 11 — Uma caixa de 15 peças sobressalentes para certo tipo de máquina contém 10 peças boas e 5 defeituosas. Se 3 unidades forem tiradas, ao acaso, da caixa, qual a probabilidade de que (a) sejam todas boas; (b) sejam todas defeituosas; (c) 2 sejam boas e 1 defeituosa; (d) pelo menos 2 sejam boas? As soluções serão expressas em termos de combinação e depois avaliadas.

(a) $$\frac{\binom{10}{3}}{\binom{15}{3}} = \frac{10!}{3!7!}\frac{3!12!}{15!} = \frac{10 \cdot 9 \cdot 8}{15 \cdot 14 \cdot 13} = \frac{24}{91}$$

(b) $$\frac{\binom{5}{3}}{\binom{15}{3}} = \frac{5!}{3!2!}\frac{3!12!}{15!} = \frac{5 \cdot 4 \cdot 3}{15 \cdot 14 \cdot 13} = \frac{2}{91}$$

(c) $$\frac{\binom{10}{2}\binom{5}{1}}{\binom{15}{3}} = \frac{10!}{2!8!}\frac{5!}{1!4!}\frac{3!12!}{15!} = \frac{10 \cdot 9}{2}\frac{5}{1}\frac{3 \cdot 2}{15 \cdot 14 \cdot 13} = \frac{45}{91}$$

(d) $$\frac{45}{91} + \frac{24}{91} = \frac{69}{91}$$

EXERCÍCIOS

Seção 2

1 — Anote todas as ocorrências possíveis do lançamento de uma moeda por 4 vezes.

2 — Uma caixa contém 1 bola vermelha, 1 preta e uma verde. As bolas devem ser retiradas desta caixa sem reposição da primeira antes da retirada da segunda. Construa um espaço amostral para este experimento semelhante ao do Quadro 1.

3 — Uma caixa contém 2 bolas pretas e 1 branca. Duas bolas devem ser retiradas desta caixa. Construa um espaço amostral para este experimento (a) usando 6 pontos, e (b) usando 3 pontos.

Seção 3

4 — Que probabilidades deveriam ser atribuídas aos pontos do espaço amostral correspondente ao experimento do problema 1?

5 — Que probabilidade você atribuiria aos pontos do espaço amostral do problema 2? Que atribuição você teria feito se a primeira bola fosse devolvida à caixa antes da segunda retirada?

6 — Que probabilidade você atribuiria aos pontos dos espaços amostrais contruídos no problema 3?

7 — Uma caixa contém 2 bolas pretas e 2 brancas. Duas bolas devem ser retiradas desta caixa. Construa um espaço amostral para este experimento e atribua probabilidades para os pontos deste espaço amostral.

8 — Se você estivesse interessado em estudar as flutuações de um mercado de ações, num determinado período de tempo, você esperaria que a freqüência relativa de altas fosse mais ou menos a mesma que a freqüência relativa de baixas? Seria possível que isso acontecesse e o mercado de ações, ainda assim, subisse constantemente, no período de um ano?

9 — Façamos $\mathbf{e}_1$, $\mathbf{e}_2$, e $\mathbf{e}_3$ simbolizar os eventos "obtenção de um número menor que 4, obtenção de número entre 4 e 6 inclusive, e obtenção de um número maior que 6", respectivamente, ao selecionarmos um número de uma tabela de números aleatórios.

(a) Construa um espaço amostral para este experimento e atribua probabilidades aos pontos.

(b) Realize o experimento 200 vezes e calcule as freqüências experimentais relativas para os três eventos a fim de verificar se seu modelo é ou não realista.

Seção 4

10 — Uma caixa contém 4 bolas vermelhas, 3 pretas, 2 verdes e 1 branca. Uma bola é retirada da caixa e depois devolvida. Qual a probabilidade de que a bola seja (a) vermelha; (b) vermelha ou preta? Simule agora este experimento por meio de números aleatórios chamando os números 0, 1, 2, 3 de vermelho, os números 4, 5, 6 de pretos, os números 7, 8 de verdes e o número 9 de branco e realize a experiência de escolher um número do quadro de números aleatórios 400 vezes. Façamos $\mathbf{A}_1$ e $\mathbf{A}_2$ significar os eventos (a) e (b) respectivamente e guarde a tabulação do número de vezes que $\mathbf{A}_1$ e $\mathbf{A}_2$ ocorreram. Observe se o modelo matemático, em confronto com estes dados experimentais, parece ser realista.

11 — Lança-se um dado honesto duas vezes. Usando o Quadro 1, calcule a probabilidade de se obter (a) um total de 4, (b) um total menor que 4, (c) um total que seja par.

12 — Uma moeda não viciada é lançada 4 vezes. Usando o modelo dos problemas 1 e 4, calcule a probabilidade de se obter (a) 4 caras, (b) 3 caras e 1 coroa, (c) pelo menos 2 caras.

13 — Para a experiência do lançamento de dois dados honestos calcule a probabilidade de que (a) a soma dos números não seja 9, (b) não apareça os números 1, 2 ou 3, (c) cada dado mostre 3 ou mais pontos, (d) os números sobre os 2 dados não sejam os mesmos, (e) exatamente um dado mostre menos que 2 pontos.

14 — Uma caixa contém 2 bolas pretas, 2 brancas e 1 verde. Duas bolas devem ser retiradas da caixa. Construa um espaço amostral bidimensional para este experimento usando 20 pontos amostrais e atribuindo iguais probabilidades para esses pontos. Usando este espaço amos-

tral e a definição (1) ou a fórmula (2), calcule a probabilidade de se obter (a) duas bolas pretas, (b) uma bola preta e uma bola branca, (c) uma bola preta e uma bola verde.

15 — Volte ao problema 14 no caso de a primeira bola retirada ser devolvida à caixa antes da retirada da segunda. Seu espaço amostral aqui terá 25 pontos.

Seção 5

16 — Escolhe-se uma carta de um baralho normal. Em tal experiência, quais dos seguintes pares de ocorrências possíveis são mutuamente exclusivos?

(a) Um paus, um rei.

(b) Um paus, uma carta vermelha.

(c) Ás de paus, uma carta preta.

(d) Ás de paus, uma carta com figura.

17 — Um espaço amostral é dividido em 3 regiões mutuamente exclusivas $\mathbf{A}_1$, $\mathbf{A}_2$ e $\mathbf{A}_3$. Se $\mathbf{P}\{\mathbf{A}_1\} = \frac{3}{6}$, $\mathbf{P}\{\mathbf{A}_2\} = \frac{2}{6}$ e $\mathbf{P}\{\mathbf{A}_3\} = \frac{1}{6}$, calcule

(a) $\mathbf{P}\{\mathbf{A}_1$ ou $\mathbf{A}_3\}$.

(b) $\mathbf{P}\{\mathbf{A}_1$ ou $\mathbf{A}_2$ ou $\mathbf{A}_3\}$.

(c) $\mathbf{P}\{$não $\mathbf{A}_1\}$.

(d) $\mathbf{P}\{$não $(\mathbf{A}_1$ ou $\mathbf{A}_2)\}$.

18 — Façamos $\mathbf{A}_1$ e $\mathbf{A}_2$ eventos associados à Figura 6. Use o espaço amostral para calcular $\mathbf{P}\{\mathbf{A}_1$ ou $\mathbf{A}_2\}$. Como $\mathbf{A}_1$ e $\mathbf{A}_2$ não são mutuamente exclusivos, a fórmula (3) não poderá ser aplicada aqui.

19 — Com a ajuda da Figura 5, mostre que a regra da adição para 2 eventos $\mathbf{A}_1$ e $\mathbf{A}_2$ que não mutuamente exclusivos, é dada pela fórmula

$$P\{A_1 \text{ ou } A_2\} = P\{A_1\} + P\{A_2\} - P\{A_1 \text{ e } A_2\}.$$

Seção 6

20 — Duas bolas devem ser retiradas de uma urna que contém 2 bolas brancas e 3 bolas pretas. Qual a probabilidade de que a primeira bola seja branca e a segunda preta?

21 — Duas bolas devem ser retiradas de uma urna que contém 2 bolas brancas, 3 pretas e 5 verdes. (a) Qual a probabilidade de que ambas as bolas sejam verdes? (b) Qual a probabilidade de que ambas as bolas sejam da mesma cor?

22 — Uma caixa contém 6 moedas, 5 das quais são honestas (não viciadas) mas a sexta tem cara de ambos os lados. Se escolhermos uma moeda da caixa e a lançarmos duas vezes, qual a probabilidade de se obter 2 caras?

23 — Um processo industrial produz 4 por cento de itens defeituosos. A experiência mostra que 25 por cento dos itens defeituosos produzidos não são percebidos pelo inspetor de qualidade. Os itens bons sempre passam, satisfatoriamente, pela inspeção. Qual a probabilidade de que, se você comprar um destes itens, seja um item defeituoso?

24 — A probabilidade de que um time de futebol vença seu próximo oponente é estimada em 0,7 se não chover, mas só em 0,5 se chover. Se os registros meteorológicos mostrarem que choveu 40 por cento das vezes, na data do jogo, nos anos passados, qual a probabilidade de que o time vença seu próximo oponente?

Seção 7

25 — Um rato é colocado num labirinto com 5 lugares onde deve escolher uma passagem correta entre duas possíveis, para conseguir alimento no fim do trajeto. (a) Se o rato fizer cada escolha, em seu primeiro trajeto, ao acaso, qual a probabilidade de que percorra o labirinto corretamente, na primeira tentativa? (b) Qual será esta probabilidade se sua probabilidade de fazer a escolha correta em cada conexão for $\frac{2}{3}$?

26 — Certo artigo manufaturado — que não poderá ser usado se tiver defeito — deve passar por duas inspeções antes que se permita seu empacotamento. A experiência mostra que um inspetor de qualidade perderá 4 por cento dos artigos defeituosos, ao passo que o segundo inspetor perderá 5 por cento dos mesmos. Os artigos não defeituosos sempre passam pela inspeção. (a) Qual a probabilidade de que um artigo defeituoso passe por ambos os inspetores? (b) Qual a probabilidade de que um artigo defeituoso passe por um inspetor mas seja apanhado pelo outro?

27 — Suponhamos que seja de 0,005 a probabilidade de que o motor de um monomotor de combate falhe na decolagem e que a taxa de falha técnica do motor de um bimotor de combate seja de 0,003. Se o bimotor não puder decolar a não ser que ambos os motores estejam operando adequadamente, que avião é mais seguro na decolagem?

28 — Certo aparelho eletrônico só opera corretamente se todos os seus componentes operarem corretamente. Há 5 componentes no aparelho. Façamos **p** significar a probabilidade de que um único componente funciona com correção e suponhamos que esta probabilidade seja a mesma para todos os componentes. Se cada componente operar independentemente dos outros, qual deve ser o tamanho de **p** para que a probabilidade de o aparelho funcionar seja, pelo menos, tão grande quanto 0,9?

29 — Os números seguintes foram obtidos de uma tabela de mortalidade baseada em 100 000 indivíduos:

Idade	Número de Vivos	Mortos por 1 000 durante esse Ano
17	94.818	7,688
18	94.089	7,727
19	93.362	7,765
20	92.637	7,805
21	91.914	7,855

Se estes números são usados para definir probabilidade de morte para o grupo de idade correspondente e se **A**, **B** e **C** indicam indivíduos de 17, 19 e 21 anos de idade, respectivamente, calcular a probabilidade de que durante o ano: (a) — **A** morra e **B** continue vivo, (b) ambos, **A** e **B**, morram, (c) ambos, **A** e **B**, continuem vivos, (d) pelo menos um de **A** e um de **B** morra, (e) pelo menos um de **A**, **B** e **C** morra.

30 — Uma organização de testes deseja avaliar uma marca particular de rádios. Selecionam-se aleatoriamente 6 rádios do estoque de rádios, e a marca será julgada satisfatória se não houver nada de errado em cada um dos 6 rádios. (a) Qual a probabilidade de que a marca venha a ser considerada satisfatória se 10 por cento dos rádios realmente são defeituosos? (b) Qual é esta probabilidade se 20 por cento são defeituosos?

31 — Sejam $\mathbf{A}_1$ e $\mathbf{A}_2$ os eventos associados à Figura 6. Use a fórmula geral do exercício 19 para calcular $\mathbf{P}\{\mathbf{A}_1 \text{ ou } \mathbf{A}_2\}$ e compare sua resposta com a do exercício 18.

Seções 9 e 10

32 — Suponhamos que 10 por cento dos possuidores de carro que têm um acidente terão tido, pelo menos, outro acidente, nos anos passados, e suponhamos ainda que um simulador de direção tenha falhado em 70 por cento de tais motoristas, mas só em 20 por cento dos que só tiveram um acidente. Se um proprietário de carro, selecionado ao acaso dentre os acidentados, faz o teste e não passa, qual a probabilidade de que ele seja do tipo dos que se acidentaram mais vezes?

33 — Suponha que haja um igual número de estudantes do sexo masculino e feminino num curso secundário e que a probabilidade seja $\frac{1}{5}$ de que um rapaz se dedique às ciências e $\frac{1}{25}$ de que uma moça o faça. Qual a probabilidade de que (a) um estudante selecionado ao acaso seja um estudante de ciências do sexo masculino? (b) um estudante escolhido ao acaso seja um estudante de ciências? (c) um estudante de ciências selecionado ao acaso seja do sexo masculino?

34 — Suponhamos um teste de aptidão colegial, destinado a separar estudantes colegiais em grupos que prometem e grupos que não prometem, ao entrarem para a faculdade. Entre os estudantes que tiveram notas satisfatórias em seu primeiro ano, 80 por cento passou no teste de aptidão. Entre os estudantes que tiveram um trabalho insatisfatório, 40 por cento passou, no teste. Supõe-se que não se use o teste para admissão, na faculdade. Se se sabe que somente 70 por cento dos alunos do primeiro ano obtiveram notas satisfatórias, qual a probabilidade de que um estudante que tenha passado no teste seja um estudante satisfatório?

35 — Um artigo manufaturado que não pode ser usado se for defeituoso, deve passar por duas inspeções antes de receber embalagem. A experiência mostra que um dos inspetores deixará passar 5 por cento dos artigos defeituosos, ao passo que o segundo inspetor deixaria passar 4 por cento de tais artigos. Se os artigos sem defeito sempre passam pela inspeção e se 10 por cento dos artigos processados são defeituosos, que porcentagem dos artigos produzidos que passam pelas duas inspeções são defeituosos?

Seções 10 e 11

36 — Quatro cartas devem ser retiradas de um baralho normal de 52 cartas. (a) Qual a probabilidade de que todas as 4 cartas sejam espadas? (b) Qual a probabilidade de que todas as 4 cartas sejam do mesmo naipe? (c) Qual a probabilidade de que nenhuma das 4 cartas seja espada?

37 — Se se retira de um baralho uma mão de pôquer de 5 cartas, qual a probabilidade de que contenha exatamente um ás?

38 — Suponhamos que seja **p** a probabilidade de que o tempo (com sol ou nublado) seja o mesmo do dia anterior. Se hoje for dia de sol, qual a probabilidade de que depois de amanhã tenhamos também um dia de sol?

39 — Empregados de certa firma são submetidos a um teste de aptidão quando empregados pela primeira vez. A experiência mostrou que dos 60 por cento que passaram no teste 80 por cento deles eram bons trabalhadores, enquanto dos 40 por cento dos que não conseguiram passar só 30 por cento foram avaliados como bons trabalhadores. Qual a probabilidade de que um empregado escolhido ao acaso seja um bom trabalhador? Use aqui a técnica da árvore.

40 — Se uma caixa contém 40 fusos bons e 10 fusos defeituosos e 8 dentre esses fusos forem selecionados, qual a probabilidade de que eles sejam todos bons? Use combinações.

41 — Ache a probabilidade de que uma mão de pôquer de 5 cartas contenha só cartas pretas, se se sabe que ela contém, pelo menos, três cartas pretas.

42 — Retira-se de um baralho de 52 cartas uma mão de bridge de 13 cartas. Use combinações para calcular a probabilidade de que a mão (a) conterá exatamente um ás, (b) conterá pelo menos um ás, (c) conterá só espadas.

43 — Um experto no conhecimento de vinhos afirma que pode facilmente diferenciar os vinhos Bordeaux dos Burgundy. Para testar sua habilidade apresenta-se a ele 6 copos de vinhos, três dos quais são três tipos diferentes de Bordeaux e os três outros são três diferentes tipos de Burgundy. São apresentados aleatoriamente. Se ele estiver informado de que três espécies são Borderaux e três Burgundy, qual a probabilidade de que, provando os vinhos, seja capaz de acertar a seleção casualmente, supondo que ele não tenha a habilidade mencionada?

Geral

44 — Uma caixa contém as 5 cartas seguintes: ás de espadas, as de paus, dois de copas, dois de ouros e três de espadas. Um ás é considerado como um. Espadas e paus são cartas pretas, copas e ouro são vermelhas. Deve-se tirar duas cartas desta caixa sem repor a primeira carta retirada antes que a segunda seja retirada. Calcule a probabilidade de que (a) ambas as cartas sejam vermelhas, (b) a primeira carta seja um ás e a segunda seja um dois, (c) ambas as cartas sejam da mesma cor, (d) uma carta seja espada e a segunda um paus, (e) seja obtido um total de 4 nas duas cartas, (f) seja obtido exatamente um ás se soubermos que ambas as cartas são pretas, (g) a primeira carta retirada seja uma espada se a segunda carta retirada for um ás de espadas.

45 — Uma caixa contém as 5 cartas seguintes: ás de espadas, ás de paus, dois de copas, dois de outros e três de espadas. Deve-se retirar duas cartas desta caixa, repondo-se a primeira carta antes da retirada da segunda. Calcule a probabilidade de que (a) ambas as cartas sejam ás, (b) o ás de espadas seja obtido, (c) pelo menos uma carta seja um ás, (d) um ás vermelho não seja obtido, (e) a soma dos números, nas duas cartas, seja menor que 4, (f) pelo menos um ás seja obtido se se sabe que nenhum cartão é um 3.

4

DISTRIBUIÇÃO DE PROBABILIDADES

1 — INTRODUÇÃO

O Capítulo 2 foi dedicado às distribuições de freqüências das amostras e à sua descrição. Este capítulo se prende às distribuições de freqüências das populações e às suas propriedades. A distribuição de freqüências de uma amostra é uma estimativa da distribuição de freqüências da população correspondente. Se o tamanho da amostra for grande, pode-se esperar que a distribuição de freqüências da amostra seja uma boa aproximação da distribuição de freqüências da população. Por exemplo, se num estudo dos pesos dos alunos internos fosse tomada uma amostra de 400 estudantes e houvesse apenas 800 estudantes internos, em toda a universidade, poder-se-ia esperar que as distribuições de freqüências da amostra e da população fossem muito semelhantes.

Na maioria dos problemas estatísticos a amostra não é suficientemente grande para determinar a distribuição da população de maneira muito precisa. Contudo, há normalmente bastante informação na amostra, juntamente com informação obtida de outras fontes, de modo a sugerir o tipo geral da distribuição da população correspondente. A experiência com várias espécies de distribuição biológicas de peso, por exemplo, mostra que elas tendem a possuir uma distribuição muito semelhante àquela vista na Figura 2 do Capítulo 2. De modo análogo, a experiência com distribuições de várias medidas lineares, como estatura, comprimento do pé e diâmetros de pistão, mostra que essas variáveis possuem distribuições muito semelhantes àquela da Figura 3 do Capítulo 2. Assim, combinando a experiência e a informação, fornecidas pela amostra, pode-se comumente convencionar a natureza geral da distribuição

da população. Esta convenção leva ao que é conhecido como distribuições de probabilidade ou distribuições teóricas.

Uma distribuição de probabilidade é um modelo matemático para a distribuição real de freqüências. Os modelos probabilísticos relacionados com jogos de azar, do Capítulo 3, são exemplos de distribuições teóricas de freqüências para certas variáveis discretas. Assim, a postulação de que cada um dos 36 casos possíveis no lançamento de dados vá ocorrer com a mesma freqüência em jogadas sucessivas conduz à distribuição teórica para os dois dados. Para o experimento de pesar 120 estudantes internos, o modelo podia ser uma distribuição de freqüências contínua, tal como a distribuição em forma de sino discutida no Capítulo 2 juntamente com a interpretação do desvio-padrão. Os modelos para as variáveis contínuas são mais difíceis de explicar que aqueles para as variáveis discretas; por conseguinte, as variáveis discretas serão consideradas em primeiro lugar.

Na discussão de distribuições amostrais e de sua teoria, costuma-se chamar uma distribuição amostral "distribuição *empírica*". A relação entre uma distribuição empírica e sua distribuição de probabilidade correspondente já foi considerada para uma variável descrita na ilustração que produziu o Quadro 1 e a Figura 1 do Capítulo 2. Uma amostra de 700 foi obtida fazendo-se 700 retiradas, com reposição, após cada retirada, de um conjunto de cartas consistente de 3 ases, 2 dois e 1 três. Dos resultados da amostragem,

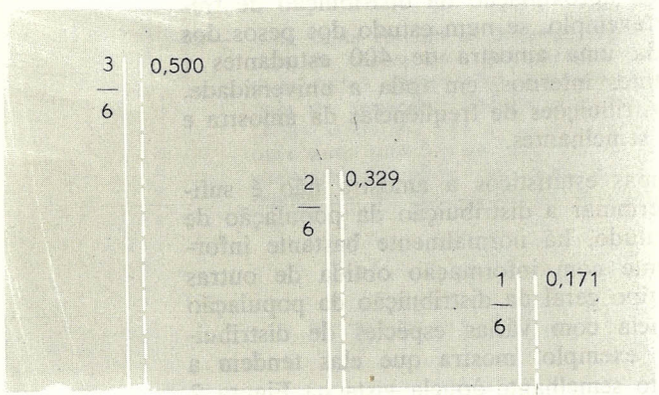


FIGURA 1. Distribuição teórica e sua aproximação empírica.

calcularam-se as freqüências relativas para cada tipo de ocorrência e, quando expressas em forma decimal, encontrou-se 0,500, 0,329 e 0,171, corrigidas para três decimais. A distribuição de probabilidade correspondente aqui é dada pelas três probabilidades $\frac{3}{6}$, $\frac{2}{6}$ e $\frac{1}{6}$. Uma comparação desta distribuição de probabilidade e sua aproximação empírica aparece na Figura 1 como um par de linhas com a linha sólida representando a teoria e a linha quebrada representando a amostra. A aproximação aqui é tão boa que, geometricamente, não se pode detectar qualquer diferença.

2 — VARIÁVEIS ALEATÓRIAS

Em experimentos do tipo repetitivo para os quais se deve construir um modelo de probabilidade, estamos geralmente interessados numa propriedade particular da ocorrência do experimento. Por exemplo, no lançamento de dois dados o interesse geralmente se concentra sobre o número total de pontos que aparece porque é tudo que realmente conta no jogo de dado. Do mesmo modo, levantando-se amostras de estudantes, o interesse deve-se concentrar nas horas semanais de estudo realizadas pelo estudante ou nas suas notas.

Como no Capítulo 2, a quantidade escolhida para estudo em tais experimentos será simbolizada pela letra x. Assim, nas ilustrações precedentes, x deveria representar a soma de pontos nos dois dados ou o número de horas semanais que um estudante gasta em estudo ou a nota do estudante. Tratando-se de distribuições de probabilidade, a variável x chama-se *variável aleatória.*

Com o fito de vermos como uma varável aleatória é introduzida num experimento simples, voltemos ao espaço amostral da experiência do lançamento de moedas dado na Figura 1 do Capítulo 3. Se x for usado para significar o número total de caras, então cada ponto do espaço amostral possuirá o valor de x mostrado logo em cima do ponto correspondente da Figura 2. Observe-se que a va-

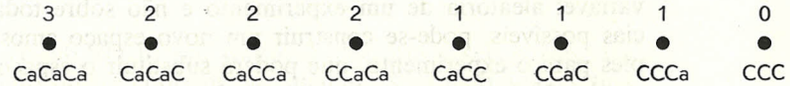


FIGURA 2. Valores de uma variável aleatória para o experimento de uma moeda.

riável aleatória x pode assumir qualquer dos valores 0, 1, 2 ou 3, mas nenhum outro valor.

Como outra ilustração, façamos x significar o número total de pontos obtidos no lançamento de dois dados não viciados. O espaço amostral deste experimento é dado pelo Quadro 1 do Capítulo 3. Este espaço amostral foi duplicado na Figura 3, mas com omis-

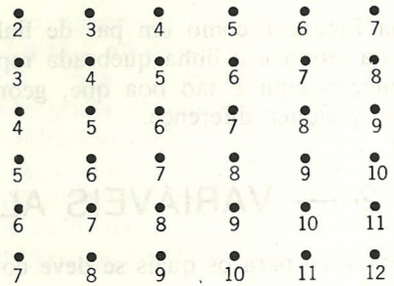


FIGURA 3. Valores de uma variável aleatória para o experimento dos dados.

são das massas associadas aos pontos. Os números ligados aos pontos são os valores da variável aleatória x para tal experimento. Observe-se também que esta variável aleatória x pode assumir qualquer dos valores 2, 3, ..., 12.

Em cada uma da ilustrações precedentes convém observar que o valor da variável x só depende do ponto amostral particular escolhido. Isto significa que x é uma função dos pontos do espaço amostral. Uma definição formal de variável aleatória seria a seguinte:

Definição. *Uma variável aleatória é uma função numérica definida sobre um espaço amostral.*

A palavra casual ou aleatória é usada para designar variáveis deste tipo que indiquem que o valor que tal variável assume num experimento depende da ocorrência do experimento que, por sua vez, depende do acaso.

Agora que o interesse está centrando-se sobre os valores de uma variável aleatória de um experimento e não sobre todas as ocorrências possíveis, pode-se construir um novo espaço amostral mais simples para o experimento, que poderá substituir o espaço amostral original. Assim, no experimento do lançamento da moeda os únicos eventos de interesse são os eventos compostos dados por $x = 0$, 1,

2 e 3. Mas, segue-se da Figura 2 e da definição (1) do Capítulo 3 que as probabilidades destes eventos compostos são dadas por $P\{0\} = \frac{1}{8}$, $P\{1\} = \frac{3}{8}$, $P\{2\} = \frac{3}{8}$ e $P\{3\} = \frac{1}{8}$, respectivamente. Estes eventos compostos podem agora ser tratados como eventos simples de um novo espaço amostral de quatro pontos com cada ponto associado ao valor da variável aleatória x. As propabilidades calculadas para estes eventos compostos são então atribuídas aos quatro pontos do novo espaço amostral. A Figura 4 mostra este novo espaço amostral com as respectivas probabilidades associadas.

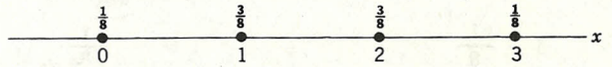


FIGURA 4. Espaço amostral para a variável aleatória do lançamento de uma moeda.

De modo semelhante, pode-se contruir um novo espaço amostral para a variável aleatória, espaço que representa o número total de pontos dos dois dados. A construção deste espaço apóia-se na Figura 3 e na definição (1) do Capítulo 3. Aqui os eventos compostos são os que correspondem à variável aleatória x e que assumem os valores 2, 3, ..., 12. As probabilidades para estes eventos compostos serão prontamente calculadas por meio da definição (1) do Capítulo 3 como sendo $\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}$ e $\frac{1}{36}$, respectivamente. Como resultado, um novo espaço amostral consistente de 11 pontos baseados nos valores da variável aleatória x pode ser construído, como mostra a Figura 5.

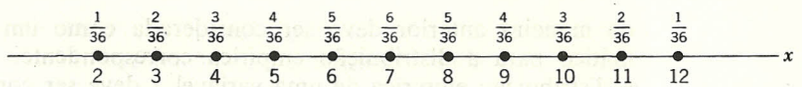


FIGURA 5. Espaço amostral para uma variável aleatória do lançamento de dados.

Depois de se introduzir uma variável aleatória e seu correspondente espaço amostral e depois de se calcular as probabilidades que devem ser atribuídas aos pontos deste espaço amostral, já se determinou a distribuição de probabilidade desejada para a variável

aleatória. Sabe-se que a distribuição de uma variável aleatória x é sempre sua distribuição de probabilidade e não sua distribuição empírica. Tal distribuição consiste em valores que x pode assumir e em probabilidades associadas a estes valores. Os gráficos das distribuições das variáveis aleatórias das Figuras 4 e 5 são dados pelas Figuras 6 e 7. Uma distribuição de probabilidade construída

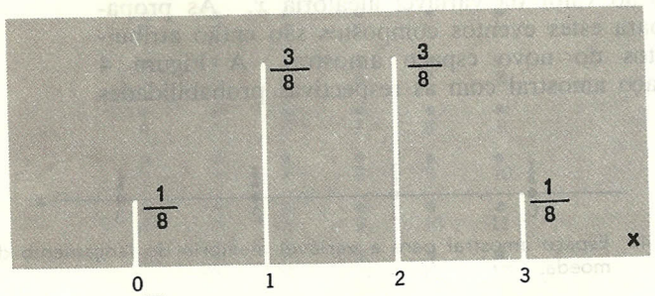


FIGURA 6. Distribuição da variável aleatória para lançamento de uma moeda.

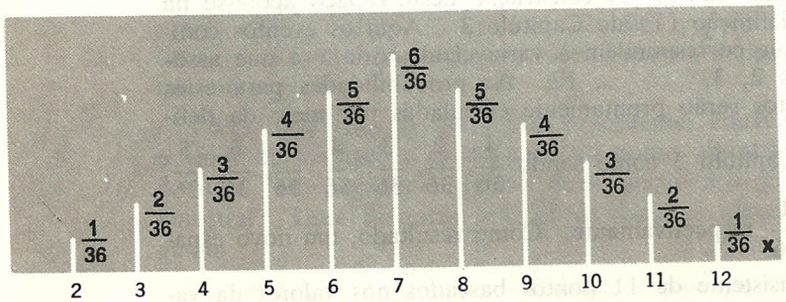


FIGURA 7. Distribuição de uma variável aleatória relacionada ao lançamento de dados.

da maneira anterior deve ser considerada como um modelo matemático para a distribuição empírica correspondente. Inversamente, a distribuição empírica de uma variável x deve ser considerada como uma aproximação da distribuição de probabilidade para x.

3 — PROPRIEDADES DAS DISTRIBUIÇÕES DE PROBABILIDADE

A discussão das distribuições empíricas do Capítulo 2 começou com uma representação geométrica por meio de histogramas e partiu

depois para uma representação aritmética parcial por meio da distribuição da média e do desvio-padrão. O mesmo processo será seguido para as distribuições de probabilidade que devem ser usadas como modelos matemáticos para distribuições empíricas. A única diferença essencial nos cálculos quando se trata de distribuições de probabilidade e de distribuições empíricas é que a primeira usa probabilidades em vez de freqüências relativas observadas. Como uma distribuição de probabilidade corresponde a uma distribuição empírica para dados classificados, usaremos as fórmulas que já foram usadas para os dados classificados.

A fórmula da média dada pela fórmula (4), do Capítulo 2, pode ser escrita assim:

$$\bar{x} = \sum_{i=1}^{k} x_i \frac{f_i}{n}$$

A média teórica correspondente, simbolizada pela letra grega μ, é obtida substituindo-se a freqüência relativa amostral f_i/n correspondente ao valor x_i pela probabilidade $P\{x_i\}$; donde,

$$\mu = \sum_{i=1}^{k} x_i P\{x_i\} \tag{1}$$

Se a amostra extraída de uma população tornar-se cada vez maior, deve-se esperar que a freqüência relativa f_i/n, de obtenção do valor x_i se aproxime de um valor fixo. Caso isso aconteça, o valor fixo deveria ser a probabilidade $P\{x_i\}$. Assim, μ poderia ser considerado como o valor do qual se esperaria que $\bar{x}$ se aproximasse, à medida que o tamanho amostral n se tornasse maior. O Quadro 1 e a Figura 1 do Capítulo 2 mostraram como as freqüências relativas se aproximam dos valores de probabilidade com o aumento do tamanho da amostra.

De modo semelhante, a fórmula para a variância de uma distribuição empírica em sua feição não modificada dada pela fórmula (6) do Capítulo 2, pode ser escrita como

$$s^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \frac{f_i}{n}$$

A variância teórica correspondente, denotada por σ^2, é obtida substituindo f_i/n por $P\{x_i\}$ e substituindo-se $\bar{x}$ por μ; donde,

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P\{x_i\} \tag{2}$$

O desvio-padrão teórico, σ, é a raiz quadrada positiva de σ^2. Não há problema aqui, como houve na definição de s^2, sobre o caso de dividir o numerador por n ou n-1, porque n não está envolvido em definições teóricas. O valor de σ^2 pode ser considerado como o valor do qual se esperaria que s^2 se aproximasse, à medida que o tamanho da amostra aumentasse.

Como uma ilustração dos cálculos envolvidos nas fórmulas (1) e (2), calcularemos a média e a variância da distribuição dada pela Figura 6. Os valores possíveis de x são $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ e $x_4 = 3$, com as probabilidades correspondentes $P\{x_1\} = \frac{1}{8}$, $P\{x_2\} = \frac{3}{8}$, $P\{x_3\} = \frac{3}{8}$ e $P\{x_4\} = \frac{1}{8}$. Usando o mesmo tipo de cálculo que o do Quadro 3, Capítulo 2, para calcularmos $\bar{x}$, o valor de μ pode ser obtido como aparece no Quadro 1.

QUADRO 1

x_i	$P\{x_i\}$	$x_i P\{x_i\}$
0	$\frac{1}{8}$	0
1	$\frac{3}{8}$	$\frac{3}{8}$
2	$\frac{3}{8}$	$\frac{6}{8}$
3	$\frac{1}{8}$	$\frac{3}{8}$
Totais	1	$\frac{12}{8} = \frac{3}{2}$

Segue-se da fórmula (1) que $\mu = \frac{3}{2}$.

Cálculos semelhantes aos usados no Quadro 4, Capítulo 2, para o cálculo de s darão o valor de σ^2. Eles aparecem no Quadro 2.

QUADRO 2

x_i	$P\{x_i\}$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 P\{x_i\}$
0	$\frac{1}{8}$	$-\frac{3}{2}$	$\frac{9}{4}$	$\frac{9}{32}$
1	$\frac{3}{8}$	$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{32}$
2	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{32}$
3	$\frac{1}{8}$	$\frac{3}{2}$	$\frac{9}{4}$	$\frac{9}{32}$
Totais	1	0		$\frac{24}{32}$

Segue-se da fórmula (2) que $\sigma^2 = \frac{24}{32} = \frac{3}{4}$.

A média e a variância de uma distribuição de probabilidade chamam-se também média de variância da variável aleatória x cuja distribuição de probabilidade é usada nos cálculos. Assim, os valores de μ e σ^2 calculados podem ser chamados de média e variância da variável aleatória x, onde x significa o número de caras obtidas, no lançamento de três moedas.

Às vezes, é mais conveniente carcularmos σ^2 por meios da fórmula seguinte, facilmente obtida modificando os parênteses e somando os termos individuais da fórmula (2).

$$\sigma^2 = \sum_{i=1}^{k} x_i^2 P\{x_i\} - \mu^2. \tag{3}$$

A variância da distribuição dada na Figura 7 será calculada para ilustrarmos o uso desta fórmula. O leitor deverá calcular o va-

lor de σ aqui, baseado na propriedade de sua distribuição normal, antes de observar que os cálculos concordam com sua intelecção e sua percepção para o valor σ. Porque a distribuição é simétrica em torno do ponto $x = 7$ a média deverá estar neste ponto de simetria; donde $\mu = 7$. Procede-se então aos cálculos como no Quadro 3.

QUADRO 3

x_i	$P\{x_i\}$	x_i^2	$x_i^2 P\{x_i\}$
2	$\frac{1}{36}$	4	$\frac{4}{36}$
3	$\frac{2}{36}$	9	$\frac{18}{36}$
4	$\frac{3}{36}$	16	$\frac{48}{36}$
5	$\frac{4}{36}$	25	$\frac{100}{36}$
6	$\frac{5}{36}$	36	$\frac{180}{36}$
7	$\frac{6}{36}$	49	$\frac{294}{36}$
8	$\frac{5}{36}$	64	$\frac{320}{36}$
9	$\frac{4}{36}$	81	$\frac{324}{36}$
10	$\frac{3}{36}$	100	$\frac{300}{36}$
11	$\frac{2}{36}$	121	$\frac{242}{36}$
12	$\frac{1}{36}$	144	$\frac{144}{36}$
Totais	1		$\frac{1974}{36}$

Segue-se da fórmula (3) que

$$\sigma^2 = \frac{1974}{36} - 49 = \frac{35}{6}$$

Portanto,

$$\sigma = \sqrt{\frac{35}{6}} \doteq 2{,}4$$

4 — VALOR ESPERADO

A média e a variância de uma variável aleatória são apenas duas de suas propriedades úteis. Podem ser tratadas como casos especiais de uma propriedade mais geral chamada *valor esperado*. Para descrevermos o significado destas duas palavras "valor esperado" consideremos um jogo no qual se lançam três moedas não viciadas e se recebe um cruzeiro por cada cara que aparecer. Quanto se esperaria ganhar se se pudesse fazer o jogo apenas uma vez? Da Figura 6 as probabilidades associadas à obtenção de 0, 1, 2 e 3 caras, no lançamento de três moedas honestas, são dadas por $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$ e $\frac{1}{8}$, respectivamente. Intuitivamente, portanto, você deveria esperar obter 0 cruzeiro $\frac{1}{8}$ das vezes, 1 cruzeiro $\frac{3}{8}$ das vezes, 2 cruzeiros $\frac{3}{8}$ das vezes e 3 cruzeiros $\frac{1}{8}$ das vezes se o jogo fosse repetido muitas vezes. Deveria esperar, portanto, o seguinte montante médio

$$\$0 \cdot \frac{1}{8} + \$1 \cdot \frac{3}{8} + \$2 \cdot \frac{3}{8} + \$3 \cdot \frac{1}{8} = \$1{,}50$$

Este montante é o que se chama comumente o montante esperado a ser ganho, se o jogo for jogado uma vez. Observe-se do Quadro 1 que este é o mesmo que o valor da média μ. O exemplo que segue é uma ilustração do conceito geral de valor esperado.

Suponhamos que a variável aleatória x deva assumir um dos valores $x_1, x_2, \ldots\ldots, x_k$ e as probabilidades associadas a estes valores sejam $P\{x_1\}, P\{x_2\}, \ldots, P\{x_k\}$, onde $\sum_{i=1}^{k} P\{x_i\} = 1$.

Então, o valor esperado desta variável aleatória é definido como sendo a quantidade

$$E[x] = \sum_{i=1}^{k} x_i P\{x_i\} \tag{4}$$

Na ilustração anterior, a variável aleatória x foi o montante de dinheiro a ser ganho no lançamento de três moedas onde os valores possíveis eram 0, 1, 2 e 3 e as probabilidades correspondentes $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$ e $\frac{1}{8}$.

Exemplo 1. Para ilustrar como se avaliam valores esperados, consideremos o seguinte problema.

Uma instituição de caridade vende bilhetes de rifa por \$ 5 cada um. Há 10 prêmios no valor de \$ 25 cada prêmio e um prêmio maior no valor de \$ 100. Se forem vendidos 200 bilhetes e você comprar um deles, qual sua expectativa em relação ao sorteio?

Aqui os valores possíveis de x são 0,25 e 100. As probabilidades correspondentes são $\frac{189}{200}$, $\frac{10}{200}$ e $\frac{1}{200}$. Segue-se da fórmula (4) que sua expectativa em unidades monetárias será

$$E[x] = \$0 \cdot \frac{189}{200} + \$25 \cdot \frac{10}{200} + \$100 \cdot \frac{1}{200} = \$1{,}75$$

Exemplo 2. Ainda como ilustração, consideremos o seguinte problema. Um lago de areia, numa quermesse escolar para angariar fundos, apresenta 25 pacotes com valor unitário de \$ 1, 5 pacotes valendo \$ 5 cada um e 1 pacote valendo \$ 10. Quanto gostaria de pagar para ter uma chance nesta pescaria, se quisesse ter um prêmio no valor daquilo que você gastou?

Os valores possíveis de x são 1, 5 e 10. As probabilidades correspondentes são $\frac{25}{31}$, $\frac{5}{31}$ e $\frac{1}{31}$. Seu valor esperado em dinheiro está dado por

$$E[x] = \$1 \cdot \frac{25}{31} + \$5 \cdot \frac{5}{31} + \$10 \cdot \frac{1}{31} = \$1{,}94$$

Portanto, você deveria pagar este montante, para ter uma chance nesta pescaria na areia.

Uma comparação da fórmula (1) com a fórmula (4) mostra que $E[x]$ nada mais é que a média μ, da variável aleatória x. Como o valor esperado de uma variável aleatória é o mesmo que sua média, pareceria não haver problema na introdução desta nova terminologia; todavia, a expectativa dá um passo a frente. Suponhamos que o jogo do lançamento das moedas seja alterado, de modo que você ganha o montante $g(x_i)$ ao invés de x_i quando se obtém o valor x_i. Por exemplo, $g(x)$ podia ser escolhido para ser a função $g(x) = x^2$. Então, o valor esperado do jogo para você seria dado pela fórmula

$$E[g(x)] = \sum_{i=1}^{k} g(x_i) P\{x_i\} \tag{5}$$

Assim, é possível falar sobre o valor esperado de qualquer função de uma variável aleatória e não da variável aleatória em si. Como ilustração, façamos $g(x) = x^2$ para o exemplo anterior do lançamento de moedas; então

$$E[g(x)] = \$0 \cdot \frac{1}{8} + \$1 \cdot \frac{3}{8} + \$4 \cdot \frac{3}{8} + \$9 \cdot \frac{1}{8} = \$3$$

Em vista deste resultado, você poderia esperar ganhar \$ 3 se os resultados finais forem os quadrados do número de caras obtidas e você tiver permissão de jogar uma vez.

Da definição (5) torna-se facilmente demonstrável que o operador E do valor esperado possui as seguintes propriedades que seguem. Nestas fórmulas c é constante.

$$E[g(x) + c] = E[g(x)] + c \tag{6}$$

e

$$E[cg(x)] = cE[g(x)] \tag{7}$$

Uma terceira propriedade mais difícil de demonstrar é dada pela fórmula

$$E[g(x) + h(y)] = E[g(x)] + E[h(y)] \tag{8}$$

Nesta última fórmula x e y são duas variáveis aleatórias quaisquer, e g e h são duas funções quaisquer dessas variáveis.

Exemplo. Para ilustrarmos o uso dessas fórmulas, suponhamos que um indivíduo, que esteja aguardando um negócio de levantamento de fundos para uma instituição de caridade, pague \$ 4 por uma chance no tabuleiro e \$ 5 por uma volta na roleta. O tabu-

leiro contém 100 furos, 20 dos quais oferecem prêmios de $ 10 e um deles oferece um prêmio de $ 100. Os 79 furos restantes são brancos. A roleta está dividida em cinco setores de igual tamanho, cada setor estando marcado por um número que vai de 0 a 4 sem repetições. A roleta paga em dólares o dobro do número no qual ela pára. Qual o valor esperado para o indivíduo que gastou $ 9?

Façamos x representar o montante a ser ganho no primeiro jogo e y representar o número onde pára a roleta. Os possíveis valores de x juntamente com as probabilidades associadas são dados por

x	0	10	100
$P\{x\}$	0,79	0,20	0,01

Os possíveis valores de y juntamente com as probabilidades associadas são dados por

y	0	1	2	3	4
$P\{y\}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

donde

$$E[x] = 0\,(0{,}79) + 10\,(0{,}20) + 100\,(0{,}01) = 3$$

e

$$E[2y] = 2E[y]$$
$$= 2\left[0\left(\frac{1}{5}\right) + 1\left(\frac{1}{5}\right) + 2\left(\frac{1}{5}\right) + 3\left(\frac{1}{5}\right) + 4\left(\frac{1}{5}\right)\right]$$
$$= 4.$$

Se $z = x + 2y$ for a variável aleatória que representa o montante total a ser ganho, em unidades monetárias, então as fórmulas anteriores afirmam que

$$E[z] = E[x] + 2E[y] = 7$$

A instituição de caridade teria, em média, um ganho de duas unidades monetárias de cada indivíduo que tivesse uma chance no jogo.

A variância da distribuição de probabilidade está convenientemente representada por um valor esperado, da mesma maneira

que a média. Isto se efetiva, escolhendo-se $g(x) = (x-\mu)^2$ na fórmula (5). Então.

$$E[(x-\mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2 P\{x_i\}$$

Uma comparação deste resultado com a fórmula (2) mostra que

$$\sigma^2 = E[(x-\mu)^2]$$

Assim, a média e a variância são vistas como casos especiais do valor esperado de uma função de uma variável aleatória.

Além de permitir a generalização da noção de valor da média de modo a incluir a média comum e a variância das distribuições de probabilidade como casos especiais, o operador do valor esperado é também um instrumento muito útil para a solução de certos tipos de problemas de tomada de decisão. Tais problemas serão estudados exaustivamente no último capítulo. Contudo, será explicado aqui um exemplo simples que indicará como se deve usar E para a solução de tais problemas.

Exemplo. Um proprietário de carro não se decide se vai comprar ou não uma apólice contra acidentes. O prêmio desta apólice é de $ 80 por ano. As estatísticas médias de sua comunidade mostram que os gastos médios com reparos de carros acidentados sobem a $ 600. As estatísticas mostram também que cerca de 10 por cento dos motoristas se acidentam, durante o ano, quer sejam eles responsáveis por isso ou não. Se este proprietário de carro tem dinheiro suficiente de modo a não ficar sobrecarregado se tiver de pagar $ 600 ou mais para reparação de seu carro, deveria adquirir a apólice contra acidentes?

A informação dada aqui pode ser disposta em forma de um quadro que levanta as perdas que podem ser esperadas, em várias possibilidades. Como a apólice de $ 100 requer um pagamento de $ 100 quando ocorre o acidente, o quadro terá a seguinte forma:

	Seguro	Nenhum acidente
Acidente	180	600
Nenhum acidente	80	0

A decisão aqui será baseada sobre os resultados do cálculo da perda esperada entre duas possíveis escolhas. Façamos E_I e E_D significar esses dois valores esperados. Os cálculos nos darão:

$$E_I = 180\left(\frac{1}{10}\right) + 80\left(\frac{9}{10}\right) = 90$$

$$E_D = 600\left(\frac{1}{10}\right) + 0\left(\frac{9}{10}\right) = 60$$

Se o proprietário do carro estiver querendo agir baseado em probabilidades que se apóiam em experiências de longo prazo, terá vantagens financeiras se não comprar a apólice de seguro. Naturalmente, este problema foi indevidamente simplificado, porque não se considerou, por exemplo, a possibilidade de se ter mais de um acidente, por ano, ou a possibilidade, mesmo se isto constitui um evento raro, de se ter um acidente muito sério. Há também vantagens psicológicas, na posse do seguro, para algumas pessoas. Se o motorista for um pouco melhor que a média dos motoristas, sua probabilidade seria menos de 1/10 e, neste caso a vantagem em não fazer seguro seria mesmo maior que $ 30.

5 — VARIÁVEIS CONTÍNUAS

Esta seção refere-se à discussão de variáveis aleatórias contínuas e sua distribuição. Tais variáveis já foram estudadas no Capítulo 2. A distinção essencial entre variável contínua e variável discreta é que a primeira envolve medida e a segunda envolve contagem. As variáveis do Capítulo 2, para as quais as técnicas de classificação dos dados foram explicadas, eram variáveis contínuas, ao passo que as variáveis do Capítulo 3, para as quais se derivaram as regras de probabilidade, eram variáveis discretas.

Os histogramas são sempre usados para representar distribuições empíricas de variáveis contínuas, sempre que os gráficos de linha forem normalmente usados para distribuições de variáveis discretas. No Capítulo 2, contudo, usamos os histogramas das Figuras 4 e 5 ainda que as variáveis, em questão, fossem discretas.

Com a finalidade de se discutir como as distribuições de probabilidade surgem em relação às variáveis contínuas, consideremos a variável contínua particular, x, que representa o diâmetro de uma barra de aço obtida da linha de produção de uma fábrica. Se as próximas 200 barras vindas da linha de produção fossem medidas, haveria 200 valores de x com que estudar a variação do diâmetro do sistema de produção. A classificação desses 200 valores de x e a feitura de histogramas ajudaria a descrever a distribuição dos

diâmetros. O histograma da Figura 8 ilustra os resultados do experimento. Se fossem medidas 400 barras, o histograma resultante teria sido cerca de duas vezes maior que o da Figura 8. Este crescimento na altura do histograma com o aumento do tamanho amostral torna difícil compararmos histogramas baseados em amostras de tamanho diferente. A dificuldade pode ser superada exigindo-se que a área do histograma seja sempre igual a 1. Pode-se conseguir isto escolhendo-se as alturas adequadas para os retângulos que compõem o histograma.

A vantagem de se usar um histograma cuja área seja igual a 1 torna-se evidente quando se calculam as freqüências relativas para diferentes conjuntos de valores de x. Assim, da Figura 8, a proporção de barras de ferro na amostra de 200, que tinham um diâmetro, em polegadas, entre 0,4705 e 0,4755, é dada pela área do retângulo para o qual esses números são os limites.

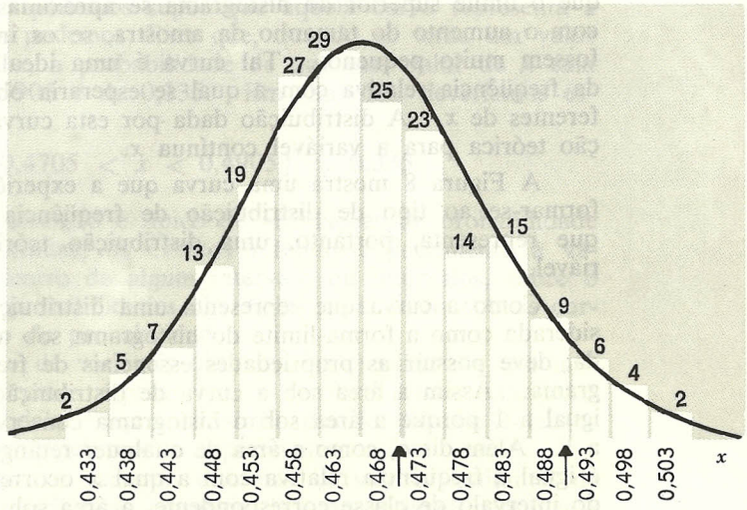


FIGURA 8. Distribuição dos diâmetros de 200 barras de aço, apresentando uma curva normal ajustada.

Estes são os limites para o intervalo cujo ponto de classe é 0,473. Esta área é, naturalmente, 23/200 = 0,115 unidades. Do mesmo modo, a proporção de barras de aço da amostra que tinham um diâmetro em polegadas entre 0,4705 e 0,4905 é dada pela soma das áreas dos retângulos que começam em 0,4705 e terminam em 0,4905 e que estão entre as duas flechas mostradas na Figura 8. Esta área é dada por (23+14+15+9)/200 = 0,305. Embora não

seja costume usar áreas para se calcular tais freqüências relativas, uma vez que elas são prontamente obtidas das freqüências observadas, é importante perceber que tais freqüências relativas podem ser representadas por áreas das partes do histograma, porque freqüências relativas teóricas serão calculadas por meio de áreas de distribuições teóricas.

Com a escolha anterior das alturas dos retângulos para se produzir uma área de 1, dever-se-ia esperar que o histograma se aproximasse de um histograma fixo com o aumento do tamanho da amostra. Por exemplo, se uma amostra de 1 000 barras de aço for retirada da linha de produção e fizermos o histograma resultante com área igual a 1, não se poderia esperar que o histograma alterasse muito sua forma se acrescentássemos mais amostra, porque o histograma já seria uma estimativa exata da distribuição dos diâmetros para o processo produtivo. Ora, se se supõe que x possa ser medido tão exatamente quanto possível de modo que o intervalo de classe se torne o menor possível, então esperar-se-ia que o limite superior do histograma se aproximasse de uma curva, com o aumento do tamanho da amostra, se os intervalos de classe fossem muito pequenos. Tal curva é uma idealização ou modelo da freqüência relativa com a qual se esperaria obter os valores diferentes de x. A distribuição dada por esta curva é uma distribuição teórica para a variável contínua x.

A Figura 8 mostra uma curva que a experiência indicou conformar-se ao tipo de distribuição de freqüência esperado aqui e que representa, portanto, uma distribuição teórica para esta variável.

Como a curva que representa uma distribuição teórica é considerada como a forma-limite do histograma sob amostragem repetida, deve possuir as propriedades essenciais de freqüência do histograma. Assim a área sob a curva de distribuição teórica deve ser igual a 1 porque a área sob o histograma conserva-se sempre igual a 1. Além disso, como a área de qualquer retângulo do histograma é igual à freqüência relativa com a qual x ocorreu entre os limites do intervalo de classe correspondente, a área sob a curva de distribuição teórica entre estes mesmos limites deveria representar a freqüência relativa esperada de x que ocorre neste intervalo. Se este mesmo raciocínio for estendido a vários intervalos vizinhos, segue-se que a área sob tal curva, entre pois valores quaisquer de x deveria representar a freqüência relativa esperada de x que ocorre dentro do intervalo determinado pelos dois valores de x.

Como ilustração consideremos, mais uma vez, os dois valores $x = 0{,}4705$ e $x = 0{,}4905$ que foram indicados pelas flechas da Figura 8. A freqüência relativa observada de x, neste intervalo, é

dada pela área dos retângulos que estão dentro do intervalo e achou-se que é 0,305. Se se supõe que a curva esboçada na Figura 8 representa uma distribuição teórica adequada, então a área debaixo desta curva entre $x = 0{,}4705$ e $x = 0{,}4905$ representará a freqüência relativa com a qual se esperaria que x ocorresse dentro deste intervalo. Esta área foi calculada, por métodos que serão explicados mais tarde, como sendo igual a 0,335.

No Capítulo 3, a probabilidade foi definida para variáveis discretas, de tal modo que foi interpretada como freqüência relativa esperada para o evento em questão. Embora a definição e as regras de probabilidade se restrinjam a variáveis discretas para facilidade de explicação, as regras valem também para eventos de variáveis contínuas. Conseqüentemente, em vista da discussão dos parágrafos anteriores, segue-se que uma curva de distribuição teórica ou de probabilidade é um curva por meio da qual se pode calcular a probabilidade de que x caia dentro de qualquer intervalo especificado sobre o eixo dos x. Assim, pela ilustração relacionada na Figura 8, enquanto podemos supor que aquela curva representa a distribuição teórica podemos dizer que, quando se toma um valor amostral único de x, a probabilidade de que esse valor de x caia entre 0,4705 e 0,4905 é de 0,335. Em símbolos, deveríamos escrever

$$\mathbf{P}\{0{,}4705 < x < 0{,}4905\} = 0{,}335$$

Este tipo de asserção é típico de afirmações de probabilidade para variáveis contínuas; os eventos referem-se geralmente à variável x que cai dentro de algum intervalo ou intervalos, sobre o eixo dos x. Tais probabilidades são dadas pelas áreas sob as curvas de distribuição de probabilidade.

Como no caso de variáveis discretas, se escolhermos um modelo realista, poderemos esperar que as probabilidades calculadas na base de tal modelo estejam próximas das freqüências relativas observadas, quando repetirmos o experimento, muitas vezes.

Como no caso das variáveis discretas, os valores limitantes de $\bar{x}$ e s^2 para uma distribuição empírica de uma variável contínua x são simbolizados pelas letras μ e σ^2. O cálculo de μ e σ^2 para distribuição de probabilidade de variável contínua, como a que foi mostrada na Figura 8, requer conhecimento de cálculo infinitesimal e, portanto, não poderá ser apresentado aqui.

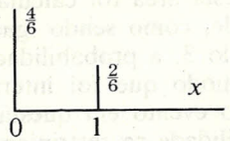
6 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Um tabuleiro tem seis furos e sabe-se que há dois prêmios de $ 1 e quatro em branco. Seja x o montante de dinhei-

ro a ser ganho. (a) 2 — Se se tiver que acertar um furo, encontre a distribuição de x e trace seu gráfico, (b) 2 — Se se tiver que acertar dois furos, ache a distribuição de x e trace o gráfico. (c) 3 — Calcule a média e o μ desvio-padrão da parte (a). Seguem-se as soluções.

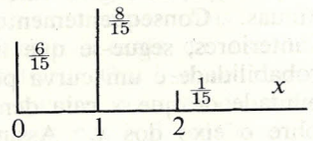
(a) $x = 0, 1$

$P\{x\} = \frac{4}{6}, \frac{2}{6}$



(b) $x = 0 \quad , \quad 1 \quad , \quad 2$

$P\{x\} = \frac{4}{6} \cdot \frac{3}{5}, \; 2 \cdot \frac{2}{6} \cdot \frac{4}{5}, \; \frac{2}{6} \cdot \frac{1}{5}$



(c) $\mu = 0 \cdot \frac{4}{6} + 1 \cdot \frac{2}{6} = \frac{1}{3}$

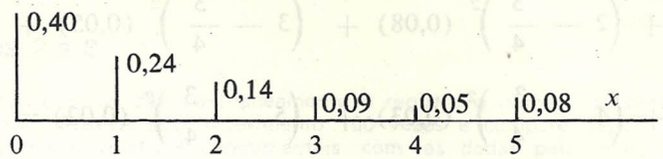
$$\sigma^2 = \left(0 - \frac{1}{3}\right)^2 \cdot \frac{4}{6} + \left(1 - \frac{1}{3}\right)^2 \cdot \frac{2}{6} = \frac{2}{9}, \; \sigma = \frac{\sqrt{2}}{3} = 0{,}47$$

Se a fórmula (3) tivesse sido usada, os cálculos teriam sido

$$\sigma^2 = 0^2 \cdot \frac{4}{6} + 1^2 \cdot \frac{2}{6} - \left(\frac{1}{3}\right)^2 = \frac{2}{6} - \frac{1}{9} = \frac{2}{9}$$

Exemplo 2. Suponhamos que seja 6 a probabilidade de que um jogador profissional de tênis, num torneio, ganhe um jogo contra seu oponente. O torneio consiste em cinco jogos. Façamos x significar o número de jogos que este jogador ganhará antes de ocorrer sua primeira derrota. (a) 2 — Ache a distribuição de x e trace o gráfico de tal distribuição. (b) 3 — Calcule a média e o desvio-padrão de x. (a) Como podemos supor que os jogos sucessivos sejam eventos independentes, a distribuição é dada por

x	0	1	2	3	4	5
$P\{x\}$	0,4	$(0,6).(0,4)$	$(0,6)^2.(0,4)$	$(0,6)^3.(0,4)$	$(0,6)^4.(0,4)$	$(0,6)^5$



(b) $\mu = 0\,(0,4) + 1\,(0,6)\,(0,4) + 2\,(0,6)^2\,(0,4) + 3\,(0,6)^3\,(0,4) + 4\,(0,6)^4\,(0,4) + 5\,(0,6)^5 = 1{,}38$

Usando a fórmula (3),

$$\sigma^2 = 0\,(0,4) + 1\,(0,6)\,(0,4) + 4\,(0,6)^2\,(0,4) + 9\,(0,6)^3\,(0,4) + 16\,(0,6)^4\,(0,4) + 25\,(0,6)^5 - (1{,}38)^2 = 2{,}46$$

donde

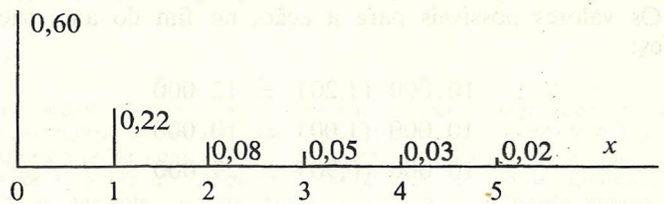
$$\sigma = 1{,}6$$

Exemplo 3. Um fabricante de pneus de automóveis conservou os registros da qualidade de seu produto e obteve o seguinte quadro de valores baseado nos últimos seis meses de produção:

Número de defeitos	0	1	2	3	4	5	≥6
Percentagem de pneus	60	22	8	5	3	2	0

Fazendo x significar o número de defeitos por pneu e tratando as percentagens da experiência como probabilidades, quando são convertidas para as proporções decimais correspondentes, (a) 2 — trace o gráfico da distribuição, (b) 3 — calcule a média e o desvio-padrão de x.

(a)



(b) $\mu = 0\,(0,60) + 1\,(0,22) + 2\,(0,08) + 3\,(0,05) + 4\,(0,03) + 5\,(0,02) = 0{,}75;$

$$\sigma^2 = \left(0 - \frac{3}{4}\right)^2 (0,60) + \left(1 - \frac{3}{4}\right)^2 (0,22) +$$

$$+ \left(2 - \frac{3}{4}\right)^2 (0{,}08) + \left(3 - \frac{3}{4}\right)^2 (0{,}05) +$$

$$+ \left(4 - \frac{3}{4}\right)^2 (0{,}03) + \left(5 - \frac{3}{4}\right)^2 (0{,}02) = 1{,}41$$

donde, $\sigma = \sqrt{1{,}41} = 1{,}2$

Exemplo 4. Um indivíduo que deseja obter uma concessão para a venda de sorvetes num acontecimento esportivo pode esperar lucrar $ 600 com a venda de sorvetes se o dia for ensolarado, mas só $ 300 se o dia estiver encoberto e $ 100 se chover. As probabilidades respectivas para esses eventos são de 0,6, 0,3 e 0,1. (a) 4 — Qual é o lucro esperado? (b) 4 — Se ele fizer um seguro no valor de $ 400 contra chuva e o custo do seguro for de $ 90, qual será o lucro esperado?

(a) $E = 600\ (0{,}6) + 300\ (0{,}3) + 100\ (0{,}1) = 460$

(b) se chover, ele perceberá tanto $ 100 como o seguro; portanto:
$E = 600\ (0{,}6) + 300\ (0{,}3) + 500\ (0{,}1) - 90 = 410$

Exemplo 5. Um investidor não sabe se investe $ 10.000 de sua herança numa ação que lhe foi recomendada ou se aplica num título de poupança que rende juros de 9 por cento. No ano seguinte a ação pode aumentar 20 por cento de valor, ou diminuir 10 por cento ou permanecer inalterada. Estima-se que as probabilidades destas três ocorrências possíveis são de 0,3, 0,4 e 0,3 respectivamente. Que deveria ele fazer se só deseja investir durante um ano e os custos de investimento são os mesmos, nos dois casos?

Os valores possíveis para a ação, no fim do ano, são os seguintes:

$$10.000\ (1{,}20) = 12.000$$
$$10.000\ (1{,}00) = 10.000$$
$$10.000\ (0{,}90) = 9.000$$

Donde, o valor esperado para este investimento é

$$12.000\ (0{,}3) + 10.000\ (0{,}4) + 9.000\ (0{,}3) = 10.300$$

O título que paga um juro de 9 por cento, por ano, produzirá o valor 10.900. Portanto, deve ser preferido.

EXERCÍCIOS

Seções 2 e 3

1 — Lance 3 moedas simultaneamente e registre o número de caras obtidas. Realize este experimento 100 vezes e compare depois as freqüências relativas experimentais com as dadas pela teoria na Figura 6.

2 — Lance um dado duas vezes simultaneamente e registre o número de pontos obtidos. Repita o experimento 100 vezes e compare suas freqüências relativas experimentais com as apresentadas pela teoria, na Figura 7.

3 — Uma caixa contém cartas que são o dois, o três e o quatro de copas. Retiram-se duas cartas da caixa, com reposição da primeira carta retirada antes da retirada da segunda. Façamos **x** significar a soma dos números obtidos, nas duas cartas. Use a técnica de enumeração de eventos para obter a distribuição da variável aleatória **x**.

4 — Resolva o problema 3, na suposição de que a primeira carta não seja devolvida à caixa.

5 — A probabilidade de que um rato percorra um labirinto com sucesso, na primeira tentativa, é de 1/4. Permite-se que 3 ratos tentem o labirinto. Façamos **x** significar o número total de trajetórias bem sucedidas. Use a técnica de enumeração de eventos para obter a distribuição da variável aleatória **x**.

6 — Calcule os valores de μ e σ para a distribuição obtida no problema 3.

7 — Calcule os valores de μ e σ para a distribuição obtida no problema 4.

8 — Calcule os valores de μ e σ para a distribuição de valores aleatórios, isto é, para $\mathbf{P}\{\mathbf{x}\} = \frac{1}{10}$, $\mathbf{x} = 0, 1, 2, \ldots, 9$.

9 — Calcule os valores de μ e σ para a distribuição dada por $\mathbf{P}\{-1\} = \frac{3}{8}$, $\mathbf{P}\{0\} = \frac{2}{8}$, $\mathbf{P}\{1\} = \frac{3}{8}$. Que percentagem da distribuição está incluída no intervalo $(\mu-\sigma, \mu+\sigma)$?

10 — Calcule a média e o desvio-padrão da distribuição do problema 5.

Seção 4

11 — Que prêmio deveria um proprietário de carro pagar por um seguro anual contra acidentes, para um carro no valor de $ 4.000, se a experiência da companhia seguradora, para carros deste valor, é dada pela tabela que segue. Suponha que se ocorrer qualquer dano será exatamente um dos danos cujo valor está listado abaixo. Não se pode cobrar danos menores que $ 500.

Dano	500	1.000	1.500	2.000	3.000
Probabilidade	0,10	0,04	0,03	0,02	0,01

12 — Você lança uma moeda três vezes. Se obtiver pelo menos duas caras, poderá lançar um dado e receberá em dólares o número de pontos que aparecer em sua face. O que você poderá esperar ganhar neste jogo, se tiver permissão de jogá-lo uma vez?

13 — Um empresário pergunta se valerá a pena fazer um seguro contra chuva, por ocasião de determinado acontecimento esportivo que ele está empresando. Se não chover espera obter $ 10.000 de renda, por ocasião da festa, mas só $ 2.000 se chover. Uma apólice de seguro de $ 7.000 lhe custará $ 3.000. Determine a probabilidade **p** de chuva, de tal modo que sua expectativa seja a mesma, faça ele o seguro ou não.

14 — Um proprietário de carro deseja vender seu carro e está estudando a possibilidade de gastar $ 50 em propaganda. Se for de 0,5 a probabilidade de que ele o venda pelo preço estipulado de $ 750 sem propaganda e de 0,9 a probabilidade de que ele o venda com propaganda, deve ou não deve fazer a propaganda? Suponha que se não o vender por $ 750 ele o venderá a um amigo por $ 650.

15 — Consideremos um jogo no qual se lança uma moeda três vezes. Recebe-se $ 5 por cada cara que apareça. Se se obtiver três coroas, será permitido lançar a moeda mais uma vez. Independentemente do sucesso, recebe-se $ 2, mas se se obtiver uma cara recebe-se $ 12 em acréscimo. O que espera ganhar, se você estiver jogando?

Geral

16 — Uma caixa contém as nove cartas seguintes: o três, o quatro e o cinco de espadas, o três e o quatro de paus, o três e o quatro de copas, e o quatro e o cinco de ouros.

(a) Se tivermos de retirar uma carta da caixa e **x** for a variável aleatória que representa o número de cartas pretas a ser obtido, ache a distribuição de **x** e trace o seu gráfico.

(b) Se tivermos de retirar uma carta e **x** representar o número sobre ela, encontre a distribuição de **x** e trace seu gráfico.

(c) Se duas cartas tiverem de ser retiradas com reposição da primeira carta antes da retirada da segunda e **x** representar o número de cartas pretas que serão obtidas, ache a distribuição de **x** e leve-a para um gráfico.

(d) Solucione o (c) no caso de não reposição da primeira carta.

17 — Que jogo escolheria se tivesse de escolher: lançar dois dados e receber em dólares a soma dos pontos ou lançar quatro moedas e receber, em dólares, o duplo do número de caras obtidas?

18 — Lança-se uma moeda três vezes. Se você obtiver menos de três caras, receberá, em dólares, o número de caras obtidas. Se conseguir três caras, poderá lançar um dado e receberá, também em dólares, o número de pontos feitos. Qual é seu valor esperado, neste jogo?

5

ALGUMAS DISTRIBUIÇÕES ESPECIAIS DE PROBABILIDADE

Neste capítulo introduziremos duas das distribuições de probabilidade mais úteis, na solução de problemas de inferência estatística. Essas distribuições, uma para variável discreta e outra para variável contínua, são indubitavelmente as duas distribuições mais comumente usadas na prática estatística.

1 — DISTRIBUIÇÃO BINOMIAL

Consideremos um experimento no qual cada uma das ocorrências possíveis pode ser classificada como resultando ou não na ocorrência de um evento A. Se resultar na ocorrência de A, será classificado como sucesso; caso contrário, como insucesso. A palavra sucesso é usada aqui como modo conveniente de descrever a ocorrência de um evento, mas não implica que tal ocorrência seja necessariamente desejada. O experimento será repetido certo número de vezes, e este número será simbolizado pela letra n. Será introduzida uma variável aleatória x que representa o número total de sucessos, isto é, ocorrências de A, obtidas nas n repetições do experimento. Uma variável aleatória deste tipo chama-se *variável binomial*.

A experiência da moeda, tantas vezes usada, pode também ser usada aqui como exemplo da variável binomial e sua distribuição. Suponhamos que a moeda seja lançada uma vez, em lugar de três. O sucesso será a obtenção de uma cara. O experimento será repetido três vezes; donde $n = 3$ aqui. A variável aleatória x representa, então, o número de caras obtidas nos três lances, como na Seção 2,

do Capítulo 4. A distribuição desta variável aleatória binomial é, portanto, dada pela Figura 6, do Capítulo 4. Está reproduzida aqui, na Figura 1.

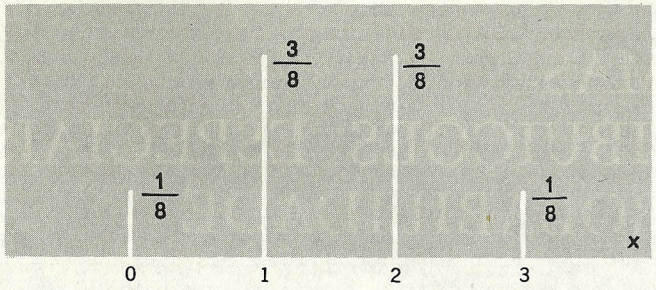


FIGURA 1. Distribuição da variável aleatória para o lançamento de uma moeda.

O experimento do lançamento de dois dados, com x sendo a soma dos pontos sobre os dois dados, não produz uma variável binomial, porque não se pode classificar cada lance de um dado como produzindo sucesso ou insucesso e x como sendo a soma destas ocorrências. Há 6 possibilidades para cada lado e não duas como seria o caso de um problema binomial.

O experimento de se retirar uma bola de uma caixa que contém três bolas vermelhas, duas pretas e uma verde também não é um experimento que conduza a uma variável binomial, porque há três possíveis ocorrências, aqui, e não duas. Mas se estivéssemos interessados somente em saber se obteremos ou não uma bola vermelha, então o problema passa a ser um problema de distribuição binomial. A variável x representaria então o número de bolas vermelhas obtidas, se realizássemos o experimento n vezes. Sempre se supõe, em problemas binomiais que dizem respeito à retirada de objetos de um recipiente qualquer, que haja reposição após cada retirada. As repetições do experimento devem ser repetições do experimento original.

Consideremos, agora, um exemplo um pouco mais complexo de variável aleatória e vejamos como se obtém sua distribuição. O experimento básico consistirá no lançamento único de um dado e o sucesso será a obtenção de um ás (um ponto). O experimento será realizado três vezes, de modo que $n = 3$, e a variável aleatória x representará o número de *uns* (1) obtidos nos três lances. Para se conseguir a distribuição desta variável aleatória pode-se proceder da mesma maneira que no experimento do lançamento das moedas

da Figura 1. Este consiste em considerar o espaço amostral original para o experimento completo e, depois, reduzi-lo a novo espaço amostral para a variável aleatória x, aplicando a definição de probabilidade ao espaço amostral original. Para cada lance de dados é preciso registrar se ocorreu sucesso ou insucesso, onde o sucesso corresponde ao aparecimento do número *um*. Se usarmos S e F para representarem sucesso e insucesso, respectivamente, haverá oito pontos no espaço amostral, como havia para três lances de moeda onde ocorria Ca ou C em cada lance. Estes pontos foram representados no Quadro 1 por meio das letras S e F que indicam as várias ocorrências possíveis. Os valores correspondentes da variável aleatória x são também mostrados no Quadro 1.

QUADRO 1

Ocorrência	SSS	SSF	SFS	FSS	SFF	FSF	FFS	FFF
Valor de x	3	2	2	2	1	1	1	0

Pode-se observar que este quadro é exatamente o mesmo que o do problema do lançamento da moeda por 3 vezes e que aparece na Figura 2, do Capítulo 4. Contudo, o cálculo de probabilidades para os vários valores de x é muito diferente. As probabilidades para estas oito ocorrências possíveis não são iguais como acontecia com o problema da moeda. As probabilidades aqui serão calculadas usando-se a regra de multiplicação das probabilidades. Devido à independência dos 3 lançamentos do dado e ao fato de que a probabilidade de um sucesso (obtenção de um 1) é agora de $\frac{1}{6}$, segue-se, por exemplo, que

$$P\{\text{SFS}\} = \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$$

Cálculos semelhantes a este fornecerão probabilidades para cada uma das oito ocorrências possíveis. Os resultados de tais cálculos aparecem no Quadro 2.

Este espaço simples de 8 pontos com as respectivas probabilidades pode agora ser reduzido ao espaço amostral para a variável aleatória x calculando-se as probabilidades dos eventos compostos correspondentes aos vários valores da variável aleatória. Isto é feito comparando-se os Quadros 1 e 2. Assim, o evento $x = 2$ compreende os eventos simples SSF, SFS, FSS. Portanto, sua probabilidade

QUADRO 2

Ocorrência	SSS	SSF	SFS	FSS	SFF	FSF	FFS	FFF
Probabilidade	$\left(\frac{1}{6}\right)^3$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$\left(\frac{5}{6}\right)^3$

é obtida somando-se as probabilidades associadas a esses três pontos. Os resultados de tais cálculos são mostrados no Quadro 3. Este quadro dá a distribuição da variável aleatória desejada. Seu gráfico

QUADRO 3

x	0	1	2	3
P {x}	$\left(\frac{5}{6}\right)^3$	$3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2$	$3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)$	$\left(\frac{1}{6}\right)^3$

aparece na Figura 2, onde as probabilidades foram expressas em forma decimal (duas decimais).

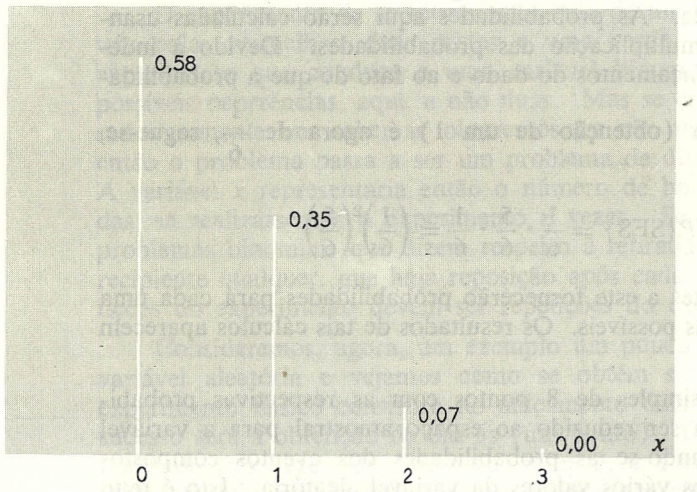


FIGURA 2. Distribuição para o número de ases no lançamento tríplice de um dado.

Em vista de tais resultados, torna-se claro que não se deve esperar três *uns*, ao se lançar três dados. Tal resultado ocorrerá, a longo prazo, cerca de 1 em $6^3 = 216$ experimentos. É claro também que não se deve aceitar uma aposta baseada na possibilidade de se ganhar a mesma quantia de modo que se obtenha pelo menos um *1* no lançamento de 3 dados. Tal resultado ocorrerá cerca de 42 por cento das vezes. Portanto se você desejar ganhar de seus amigos ingênuos, faça-os apostar que eles não conseguirão sequer um *1* no lançamento de três dados.

A técnica empregada nas duas ilustrações anteriores pode ser usada para qualquer problema binomial. Constrói-se primeiro o espaço amostral para o experimento completo. Se houver, digamos, cinco repetições do evento básico, haverá $2^5 = 32$ pontos, no espaço amostral. Calcula-se então a probabilidade para cada ponto. Depois associa-se o valor adequado da variável x a cada ponto do espaço amostral. Somando-se depois as probabilidades destes pontos que correspondem a um valor particular de x, obtêm-se as probabilidades dos vários valores de x. Estas probabilidades dão a distribuição de probabilidade desejada da variável aleatória binomial x.

Devido à dificuldade de se efetivar os cálculos precedentes toda vez que surgir um problema binomial, será conveniente ter uma fórmula aplicável a todos estes problemas. Para tanto vamos considerar um problema binomial genérico. Deve-se realizar um experimento para o qual a ocorrência possa sempre ser classificada como sucesso ou como insucesso. Supõe-se determinada probabilidade que produzirá um sucesso e será simbolizada pela letra p. A probabilidade correspondente ao insucesso será simbolizada por q; donde $p + q = 1$. O experimento deve ser realizado n vezes. O número de sucessos obtidos nas n repetições do experimento é simbolizado pela letra x. O problema então é calcular as probabilidades dos vários valores possíveis da variável aleatória x. Como ficou dito antes, tais cálculos podem ser efetuados sistematicamente para qualquer problema dado e podem também ser realizados aqui para este problema geral se pudermos utilizar da seção 7 do Capítulo 3. Isto está feito na seção 5 deste capítulo para benefício daqueles que tiverem tempo de estudar o material precedente; portanto, só será dado aqui o resultado dos cálculos. O símbolo $P\{x\}$ será usado para significar a probabilidade para um valor típico desta variável binomial genérica. Estas probabilidades, que definem o que é conhecido como distribuição binomial, são dadas pela seguinte fórmula

(1) **Distribuição binomial** $P\{x\} = \dfrac{n!}{x!(n-x)!}p^x q^{n-x}$

Como foi explicado na seção 11 do Capítulo 3, o símbolo *n!* é conhecido como "*n* fatorial". Simboliza o produto de todos os inteiros de 1 a *n*. Assim, 5! = 1.2.3.4.5. Como foi assinalado aqui, o valor de 0! é definido como sendo 1. Para valores grandes de *n* e *x*, os cálculos de $n!/x!\ (n\text{-}x)!$ tornam-se muito trabalhosos; conseqüêntemente, foi feita uma tabela de valores desta expressão para os vários valores de *n* e *x*. É a Tabela II, do apêndice. Esta tabela trabalha com valores de *n* de 2 a 20 e de *x* de 2 a $n/2$ ou mais. Não é necessário tabelarmos valores para números de *x* maiores, porque, se *x* for substituído por *n-x*, nesta expressão ter-se-á o mesmo valor. Usa-se, pois, este artifício para tratarmos valores de *x* maiores que $n/2$. Embora esta tabela nos permita obter, com rapidez, probabilidades binomiais, não será usada nos problemas que seguem, porque é desejável que o aluno tome conhecimento da fórmula (1) efetuando os cálculos requeridos quando *n* for bem pequeno.

É comum falar-se dos *n* experimentos como *n* tentativas independentes de um experimento para o qual *p* é a probabilidade de sucesso de uma única tentativa. Nesta linguagem $P\{x\}$ é a probabilidade de se obter *x* sucessos em *n* tentativas independentes de um experimento para o qual *p* é a probabilidade de sucesso para uma única tentativa.

As duas distribuições de freqüência mostradas nas Figuras 1 e 2 são casos especiais da distribuição binomial geral dada pela fórmula (1). Para o problema da moeda, $n = 3$ e $p = q = \dfrac{1}{2}$. Para o problema do dado, $n = 3$, $p = \dfrac{1}{6}$ e $q = \dfrac{5}{6}$. Os valores de $P\{x\}$ dados na Figura 1 e Figura 2 devem ser calculados por meio da fórmula (1) para nos familiarizarmos com seu uso. Como ilustração deste cálculo, se o valor de $P\{3\}$ for desejado para o problema do dado, a substituição dos valores próprios na fórmula (1) fornecerá

$$P\{3\} = \frac{3!}{3!\,0!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0$$

Como 0! = 1 e como pela álgebra qualquer número elevado à potência 0 é igual a 1, segue-se que $\left(\dfrac{5}{6}\right)^0 = 1$, e daí que

$$P\{3\} = \left(\frac{1}{6}\right)^3$$

Embora os problemas usados na apresentação da distribuição binomial se refiram a jogos de azar, a distribuição binomial é muito útil para resolver certos tipos de problemas práticos. Tais problemas são resolvidos nos próximos capítulos. Neste ínterim, discutiremos alguns problemas simples que só requerem cálculos fáceis que envolvem a fórmula (1).

Exemplo 1. A probabilidade de que certos pais com certo tipo de olhos azuis-escuros tenham filhos com olhos azuis é de $\dfrac{1}{4}$. Se houver 6 filhos na família, qual a probabilidade de que, pelo menos, metade das crianças tenha olhos azuis?

Na solução do problema, as 6 crianças da família serão tratadas como 6 tentativas independentes de um experimento para o qual a probabilidade de sucesso de uma única tentativa é de $\dfrac{1}{4}$. Assim $n = 6$ e $p = \dfrac{1}{4}$ aqui. É necessário calcular $P\{3\}$, $P\{4\}$, $P\{5\}$ e $P\{6\}$ e somar, porque estas probabilidades correspondem a maneiras mutuamente exclusivas de ocorrência do evento desejado. Pela fórmula (1),

$$P\{3\} = \frac{6!}{3!\,3!}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^3 = \frac{540}{4096}$$

$$P\{4\} = \frac{6!}{4!\,2!}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^2 = \frac{135}{4096}$$

$$P\{5\} = \frac{6!}{5!\,1!}\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^1 = \frac{18}{4096}$$

$$P\{6\} = \frac{6!}{6!\,0!}\left(\frac{1}{4}\right)^6\left(\frac{3}{4}\right)^0 = \frac{1}{4096}$$

A probabilidade de se conseguir, pelo menos, 3 sucessos é obtida somando-se estas probabilidades; conseqüentemente, escrevendo-se $x \geq 3$ para representar, pelo menos três sucessos obtém-se

$$P\{x \geq 3\} = \frac{694}{4096} = 0{,}169$$

Este resultado mostra que há uma chance muito pequena de que tal família tenha tantas crianças de olhos azuis. Só 17 dentre 100 dessas famílias terão, pelo menos, a metade dos filhos com olhos azuis.

Exemplo 2. Um fabricante de peças de automóveis garante que uma caixa de suas peças conterá, no máximo, dois itens defeituosos. Se a caixa contém 20 peças e a experiência mostra que este processo de fabricação produz dois por cento de itens defeituosos, qual a probabilidade de que uma caixa satisfaça a garantia?

Este problema pode ser considerado como um problema de distribuição binomial para o qual $n = 20$ e $p = 0{,}02$. Uma caixa satisfará a garantia se o número de peças defeituosas for 0,1 ou 2. Por meio da fórmula (1) as probabilidades destes três eventos são dadas por

$$P\{0\} = \frac{20!}{0!\,20!}(0{,}02)^0(0{,}98)^{20} = (0{,}98)^{20} = 0{,}668$$

$$P\{1\} = \frac{20!}{1!\,19!}(0{,}02)^1(0{,}98)^{19} = 20\,(0{,}02)\,(0{,}98)^{19} = 0{,}273$$

$$P\{2\} = \frac{20!}{2!18!}(0{,}02)^2(0{,}98)^{18} = 190\,(0{,}02)^2\,(0{,}98)^{18} = 0{,}053$$

Como se trata de eventos mutuamente exclusivos, a probabilidade de que haja, no máximo, duas peças defeituosas expressas por $x \leq 2$, será a soma destas probabilidades; donde a resposta desejada será

$$P\{x \leq 2\} = 0{,}994$$

Este resultado mostra que a garantia do fabricante quase sempre será satisfeita.

Exemplo 3. Como ilustração final, consideremos o seguinte problema que nos diz se vale ou não vale a pena a adivinhação num exame. Suponhamos um exame consistente num grande número de questões do tipo múltipla-escolha, com cada questão tendo 5 respostas possíveis, mas só uma das 5 sendo a resposta correta. Se o estudante receber 3 pontos por cada resposta correta e — 1 (menos um) ponto por cada resposta incorreta, e se para cada 10 questões a probabilidade de se adivinhar a resposta correta for de $\frac{1}{3}$, qual a probabilidade de se obter uma nota total positiva nestas 10 questões?

Se x significa o número de questões respondidas corretamente teremos uma nota positiva se $3x > 10 - x$, porque o primeiro membro desta inequação dá o número total de pontos positivos alcançados e o segundo membro o número total de pontos perdidos. Esta desigualdade será satisfeita se $x > \frac{10}{4}$ o que implica que, pelo menos, devemos obter três respostas corretas para alcançarmos nota positiva. Portanto, a probabilidade desejada é dada por

$$P\{x \geq 3\} = 1 - \sum_{x=0}^{2} \frac{10!}{x!(10-x)!}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{10-x}$$

$$= 1 - \left[\left(\frac{2}{3}\right)^{10} + 10\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^{9} + 45\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{8}\right]$$

$$= 0{,}70$$

Assim, se sua probabilidade for de 1/3, ele terá excelente oportunidade de ganhar pontos, usando apenas a adivinhação. Se ele não souber nada sobre a matéria e escolher uma dentre as cinco respostas possíveis, ao acaso, sua probabilidade seria, naturalmente, de $\frac{1}{5}$ em cada questão. Supõe-se aqui, contudo, que ele tenha conhecimento da matéria, o suficiente para descartar 2 das cinco possibilidades como sendo claramente incorretas e que escolha dentre as 3 restantes. Se não tiver tal conhecimento, de modo que sua possibilidade passasse a ser 1/5, então cálculos semelhantes mostrariam que não vale a pena a adivinhação.

Mesmo com a ajuda da Tabela II, os cálculos das probabilidades binomiais são, freqüentemente, trabalhosos quando se requer a soma de várias probabilidades sem calculadora eletrônica.

Em tais casos, será útil a Tabela III do apêndice, se o valor de n não exceder 10. Esta tabela apresenta a soma das probabilidades da cauda direita, na forma:

$$P\{x \geq x_0\} = \sum_{x=x_0}^{n} \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

Como os valores disponíveis de p começam em 0,05 e procedem por passos de 0,05 até o valor de 0,50, será necessário fazer interpolações para outros valores de $p < 0{,}50$. Para valores de $p > 0{,}50$, usa-se $q = 1 - p$ em lugar de p e $n\text{-}x_0 + 1$ no lugar do x_0 da tabela e subtrai-se o resultado de 1 para se obter $P\{x \geq x_0\}$. Para $n > 10$ será necessário recorrer a probabilidades aproximadas. Apresentaremos um método para isto na seção 4.

Para ilustrarmos o uso da Tabela III, consideremos sua aplicação ao último dos três exemplos anteriores. Os valores requeridos da Tabela III são $n = 10, p = \frac{1}{3} = 0{,}33$ e $x_0 = 3$. Fazendo interpo-

lação para $p = 0{,}33$ teremos $P\{x \geq 3\} = 0{,}70$, que foi a resposta obtida antes.

2 — PROPRIEDADES DA DISTRIBUIÇÃO BINOMIAL

No Capítulo 4 apresentamos fórmulas gerais para cálculo da média e variância de variáveis aleatórias discretas. Estas fórmulas podem também ser usadas para obtermos a média e a variância de qualquer variável binomial. Os cálculos que produziram os Quadros 1 e 2 do Capítulo 4 são os necessários para a variável binomial particular para a qual $p = \frac{1}{2}$ e $n = 3$.

Para nos familiarizarmos mais com tais cálculos, consideremos o cálculo da média e da variância da variável binomial dado pelo Quadro 3, cujo gráfico aparece na Figura 2. Aqui $p = \frac{1}{6}$, $n = 3$ e os valores possíveis de x são $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ e $x_4 = 3$. Os valores correspondentes de $P\{x_i\}$ são dados no Quadro 3.

O cálculo de μ, baseado nestes valores, aparece no Quadro 4.

QUADRO 4

x	P{x}	xP{x}
0	$\frac{125}{216}$	0
1	$\frac{75}{216}$	$\frac{75}{216}$
2	$\frac{15}{216}$	$\frac{30}{216}$
3	$\frac{1}{216}$	$\frac{3}{216}$
		$\frac{108}{216}$

Destes cálculos segue-se que

$$\mu = \frac{108}{216} = \frac{1}{2}$$

O cálculo de σ aparece no Quadro 5

QUADRO 5

x	P{x}	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 P\{x\}$
0	$\frac{125}{216}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{125}{864}$
1	$\frac{75}{216}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{75}{864}$
2	$\frac{15}{216}$	$\frac{3}{2}$	$\frac{9}{4}$	$\frac{135}{864}$
3	$\frac{1}{216}$	$\frac{5}{2}$	$\frac{25}{4}$	$\frac{25}{864}$
				$\frac{360}{864}$

Destes cálculos segue-se que

$$\sigma = \sqrt{\frac{360}{864}} = \sqrt{\frac{5}{12}}$$

É possível empregar alguns truques algébricos para levar a cabo cálculos semelhantes para a distribuição binomial geral dada pela fórmula (1). Tais cálculos fornecem fórmulas para a média e o desvio-padrão que podem ser usadas para todos os problemas binomiais. Como tais cálculos são um tanto complicados, não serão efetuados aqui, mas as fórmulas resultantes serão dadas e usadas para a solução de problemas binomiais. Estas fórmulas, muito simples, são

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

A vantagem de termos tais fórmulas claras torna-se evidente quando as aplicamos ao problema computacional que acabou de ser mencionado. Para a distribuição do Quadro 3, $n = 3$ e $p = \frac{1}{6}$. Portanto da fórmula (2) temos:

$$\mu = 3 \cdot \frac{1}{6} = \frac{1}{2}$$

$$\sigma = \sqrt{3 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \sqrt{\frac{5}{12}}$$

Tivemos grande poupança de tempo se compararmos o tempo despendido agora com o despendido no cálculo de tais quantidades por meio das fórmulas gerais dos Quadros 4 e 5.

3 — DISTRIBUIÇÃO NORMAL

A distribuição binomial da seção 1 é a distribuição teórica de freqüência mais útil para variáveis discretas. A distribuição a ser estudada nesta seção é a distribuição teórica de freqüências mais útil para variáveis contínuas.

O histograma da Figura 8, Capítulo 4, é típico de muitas distribuições encontradas na natureza e na indústria. Tais distribuições são bastante simétricas, extinguem-se muito rapidamente para as extremidades e possuem uma forma muito semelhante à do sino. Uma distribuição teórica, que se tem provado de muita utilidade para distribuições como estas e que presentemente será vista como muito importante também para os outros casos, é chamada *distribuição normal*. A curva esboçada na Figura 8, Capítulo 4 é o gráfico de uma distribuição normal particular. O gráfico de uma distribuição normal geral é apresentado na Figura 3. Embora uma distribuição normal seja definida pela equação da sua curva, esta equação não será usada explicitamente nos capítulos subseqüentes e por isso não será escrita aqui. A própria curva pode ser imaginada como definindo a distribuição.

Muito já se falou sobre a curva normal associada à distribuição de notas em grandes turmas de certos cursos. A graduação das notas com base na curva normal admite que a distribuição do rendimento mental dos estudantes deve ser semelhante à distribuição de muitas de suas características físicas, as quais se sabe serem distribuídas de maneira aproximadamente normal.

Um instrutor usando tal esquema de dar notas também admite comumente que os estudantes no seu curso tenham os requisitos convenientes e estudem o número de horas desejado. Eis um otimista! A dificuldade com este sistema é que ele atribui notas numa base relativa ao invés de absoluta. Dessa forma, uma classe prendada de maneira invulgar que trabalhe muito terá uma mesma distribuição de notas que uma classe de alunos negligentes. Certamente que a resposta a esta crítica é que os estudantes de ano para ano diferem muito pouco e, por conseguinte, é altamente improvável que uma turma grande seja constituída de estudiosos brilhantes ou de vadios pouco inteligentes. Se você ficou desgostoso com a curva normal com base nas notas que ela lhe atribui, compreenda que ela realmente não é culpada e imbua-se de idealismo e entusiasmo, o suficiente para aproximar-se do seu estudo sem prevenção. É uma curva excepcionalmente útil em muitos ramos da aplicação, bem afastados dos problemas escolares.

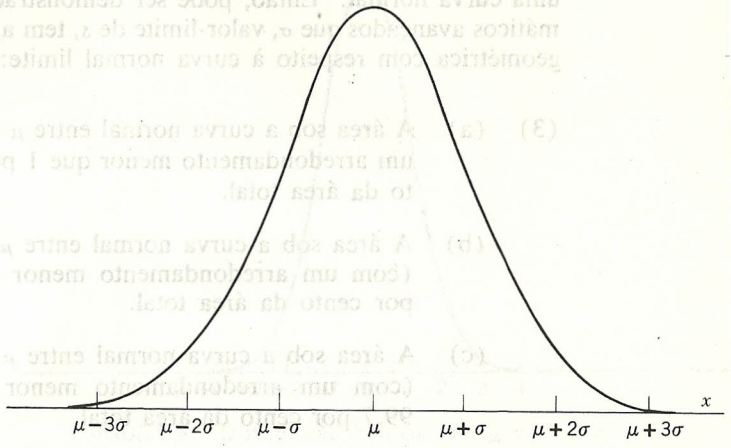


FIGURA 3. Distribuição normal típica.

Vimos no Capítulo 2 que a média aritmética de uma distribuição representa o ponto do eixo dos x no qual uma folha de metal com a forma do histograma da distribuição se equilibrará na extremidade de uma faca. Esta propriedade geométrica da média aritmética deixa claro que quando um histograma é simétrico com respeito a um eixo vertical ele deve estar localizado no ponto de simetria no eixo dos x. Isto também é verdadeiro para o valor-limite da média quando se aumenta indefinidamente o tamanho da amostra e se torna o intervalo de classe progressivamente menor.

Isso explica por que no ponto de simetria do eixo dos x na Figura 3 foi escrito o μ.

Suponhamos, agora, que a forma-limite do histograma para uma distribuição de freqüências, no sentido anteriormente descrito, seja uma curva normal. Então, pode ser demonstrado por métodos matemáticos avançados que σ, valor-limite de s, tem a seguinte interpretação geométrica com respeito à curva normal limite:

(3) (a) A área sob a curva normal entre $\mu - \sigma$ e $\mu + \sigma$ é (com um arredondamento menor que 1 por cento) 68 por cento da área total.

(b) A área sob a curva normal entre $\mu - 2\sigma$ e $\mu + 2\sigma$ é (com um arredondamento menor que 1 por cento) 95 por cento da área total.

(c) A área sob a curva normal entre $\mu - 3\sigma$ e $\mu + 3\sigma$ é (com um arredondamento menor que 0,1 por cento) 99,7 por cento da área total.

O eixo da Figura 3 foi marcado em unidade de σ, começando da média μ. Decorre do que foi observado acima que não há quase nenhuma área sob a curva a uma distância maior que 3 σ unidades a partir de μ; contudo, a sua equação mostra que ela realmente se estende de $-\infty$ a $+\infty$.

As duas primeiras propriedades anteriores foram usadas para dar significação ao s do Capítulo 2. Foi visto então que as duas percentagens são aproximadamente corretas para histogramas cujas formas se assemelham à curva normal.

Uma propriedade interessante da curva normal é que a sua localização e forma ficam completamente determinadas pelos seus valores de μ e σ. O valor de μ, é claro, centra a curva, enquanto o valor de σ determina a extensão do espalhamento. Visto que todas as curvas normais representativas das distribuições teóricas de freqüências têm uma área total de 1, quando aumenta a curva deve decrescer em altura e espalhar-se.

Isto é ilustrado nas Figuras 4 e 5 que esboçam duas curvas normais com a mesma média, a saber 0, e desvios-padrão de 1 e 3, respectivamente. O fato de a forma da curva normal ser completamente determinada pelo seu desvio-padrão permite que se reduzam todas as curvas normais a uma curva-padrão, por simples mudança de variável. Por exemplo, a curva da Figura 4 pode assemelhar-se à curva da Figura 5 se alterarmos a escala sobre o eixo dos x de modo que uma unidade sobre o eixo da Figura 4 represente três

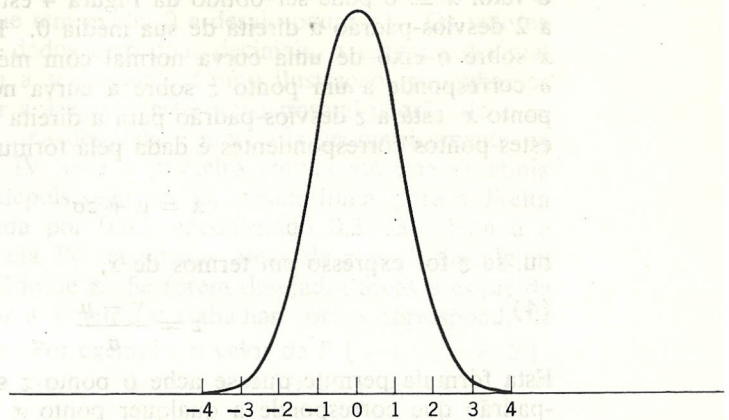


FIGURA 4. Distribuição normal com $\mu = 0$ e $\sigma = 1$.

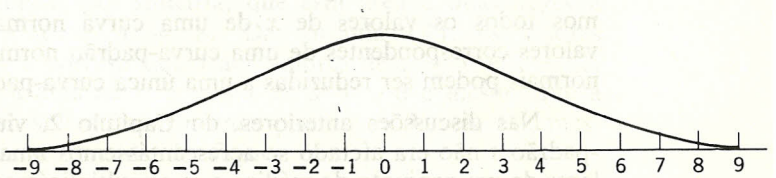


FIGURA 5. Distribuição normal com $\mu = 0$ e $\sigma = 3$.

unidades sobre o eixo da Figura 5. Isto corresponde, grosseiramente, a tomar a curva da Figura 4, tratar a ela bem como à área abaixo dela como se fossem de borracha elástica e estendê-las até que se tornem três vezes mais compridas, conservando, porém, a mesma área. Inversamente, a Figura 5 poderia tornar-se a Figura 4 se a comprimíssemos até que ela se tornasse um terço de seu valor atual. Isto supõe que as longas caudas destas curvas sejam ignoradas. Como a curva mais simples com que trabalhar é a que tem média 0 (zero) e desvio-padrão 1, outras curvas normais são reduzidas a esta curva-padrão quando há necessidade de redução. Ora, para qualquer ponto sobre o eixo x de uma curva normal existe um ponto sobre o eixo x de curva normal-padrão, e seu valor pode ser determinado estabelecendo-se por quantos desvios-padrão ela se afasta do ponto médio da curva. Assim o ponto $x = 6$ da Figura 5 corresponde ao

ponto $x = 2$ da curva normal-padrão dada pela Figura 4; portanto o valor $x = 6$ pode ser obtido da Figura 4 estabelecendo-se que está a 2 desvios-padrão à direita de sua média 0. Em geral, se um ponto x sobre o eixo de uma curva normal com média μ e desvio-padrão σ corresponde a um ponto z sobre a curva normal-padrão, então o ponto x está a z desvios-padrão para a direita de μ. A relação entre estes pontos correspondentes é dada pela fórmula

$$x = \mu + z\sigma$$

ou, se z for expresso em termos de x,

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

Esta fórmula permite que se ache o ponto z sobre a curva normal-padrão que corresponde a qualquer ponto x sobre a curva normal não-padrão. Assim, o ponto x = 4 da Figura 5 corresponde ao ponto $x = (4-0)/3 = 1\frac{1}{3}$ da Figura 4. Com tal recurso de expressarmos todos os valores de x de uma curva normal, em termos dos valores correspondentes de uma curva-padrão normal, todas as curvas normais podem ser reduzidas a uma única curva-padrão.

Nas discussões anteriores, do Capítulo 2, viu-se que o desvio-padrão s não era afetado se acrescentássemos uma constante aos valores de um conjunto de médias e se a multiplicássemos por c, se cada uma das medidas também o fosse. Esta mesma propriedade valerá para uma variável aleatória x e seu desvio-padrão teórico σ. Por outras palavras, a variável x-c terá o mesmo desvio-padrão que a variável x, mas o desvio-padrão para cx será c vezes maior para x. Em vista destas propriedades, o desvio-padrão de $(x-\mu)/\sigma$ será $1/\sigma$ vezes o desvio-padrão de $x-\mu$, ou de x; portanto, se σ é o desvio-padrão de x, o desvio-padrão de $z = (x-\mu)/\sigma$ deve ser 1. Como a média de x é μ, subtraindo-se μ de x obteremos uma variável $x-\mu$ cuja média é zero. A média da variável $z = (x-\mu)/\sigma$ será portanto também 0, porque, multiplicando-se uma variável por uma constante, multiplica-se a média por esta constante, e multiplicando-se 0 por $1/\sigma$ ainda teremos 0. Assim, a variável $z = (x-\mu)/\sigma$ possuirá a média 0 e o desvio-padrão 1. A mudança da variável dada pela fórmula (4) alterará, portanto, qualquer variável x com média 0 e desvio-padrão 1. Isto é verdade quer a variável x seja uma variável normal ou não. Diz-se que uma variável x que foi mudada para uma variável z, por meio da fórmula (4), está medida em *unidades-padrão* após efetuada a mudança.

A Tabela IV do apêndice é uma tabela destinada a mostrar a área para a variável z, sob qualquer porção da curva normal, isto é, para a curva normal que tem média 0 e desvio-padrão 1. Os valores de z, nesta tabela, são dados com duas decimais, a segunda decimal determinando a coluna a ser usada. Como ilustração, suponhamos que se queira encontrar a área sob esta curva normal-padrão de $z = 0$ a $z = 1{,}00$. A área desejada é mostrada geometricamente na Figura 6. Na Tabela IV lê-se a primeira coluna até que se atinja o valor de 1,0 para z; depois segue-se, na mesma linha, para a direita até a coluna encabeçada por 0,00, encontrando 0,3413. Esta é a área desejada. A Tabela IV dá apenas áreas de $z = 0$ até algum valor positivo especificado de z. Se forem desejadas áreas à esquerda de $z = 0$, deve-se usar a simetria e trabalhar com a correspondente metade direita da curva. Por exemplo, o valor de $P\{-1 < z < 2\}$ é dado pela soma de $P\{0 < z < 2\}$ e $P\{0 < z < 1\}$.

Suponhamos agora que se deseje encontrar a área sob a parte de uma curva normal com média μ e desvio-padrão σ. Por exemplo, suponhamos que se deseje encontrar a área de $x = \mu - \sigma$ a $x = \mu + \sigma$ na Figura 3. Segue-se, por simetria, que esta área é duas vezes a área de $x = \mu$ a $x = \mu + \sigma$. Da fórmula (4), os valores $x = \mu$ e $x = \mu + \sigma$ correspondem aos valores de $z = 0$ e $x = 1$; conseqüentemente, a área de $x = \mu$ a $x = \mu + \sigma$ é a mesma que a correspondente área sob a curva normal-padrão da Tabela IV, mostrada como a área sombreada da Figura 6.

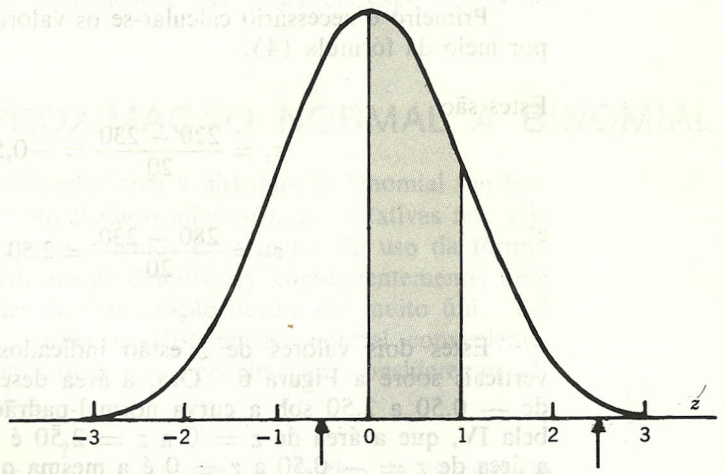


FIGURA 6. Distribuição normal-padrão.

Como se sabe que esta área é 0,3413, a área de $x = \mu - \sigma$ a $x = \mu + \sigma$ é duas vezes este número, ou seja, 0,6826. Este, naturalmente, é o número que deu origem à propriedade (a) da distribuição normal, na fórmula (3).

Exemplo 1. Como ilustração ao uso da Tabela IV, consideremos o seguinte problema. Suponhamos que a altura dos estudantes do sexo masculino de certa universidade seja uma variável normal, simbolizada por x, com 69 polegadas de média e desvio-padrão 3. Que percentagem de estudantes é maior que 6 pés? Obtém-se isto calculando a quantidade $P\{x > 72\}$.

Escolha o valor $x = 72$. Então como $\mu = 69$ e $\sigma = 3$, o valor correspondente de z é dado por

$$z = \frac{72 - 69}{3} = 1$$

Da Tabela IV, a área sob a curva normal-padrão entre $z = 0$ e $z = 1$ é 0,3413. Como a área sob a metade direita desta curva é 0,5, segue-se que a área para a direita de $z = 1$ é igual a $0{,}5000 + - 0{,}3413 = 0{,}1587$. Donde se conclui que cerca de 16 por cento dos estudantes terão um valor de z que excede a 1 e, portanto, uma altura que excede a 6 pés de altura.

Exemplo 2. Vejamos outro exemplo que esclarece o uso da Tabela IV. Suponhamos que se queira encontrar a área entre $x = 220$ e $x = 280$ para x com distribuição normal tendo $\mu = 230$ e $\sigma = 20$. A área desejada é mostrada na Figura 7.

Primeiro é necessário calcular-se os valores de z correspondentes por meio da fórmula (4).

Estes são

$$z_1 = \frac{220 - 230}{20} = -0{,}50$$

e

$$z_2 = \frac{280 - 230}{20} = 2{,}50$$

Estes dois valores de z estão indicados por meio de flechas verticais sobre a Figura 6. Ora, a área desejada é dada pela área de — 0,50 a 2,50 sob a curva normal-padrão. Sabemos, pela Tabela IV, que a área de $z = 0$ a $z = 2{,}50$ é 0,4938. Por simetria, a área de $z = -0{,}50$ a $z = 0$ é a mesma que a de $z = 0$ a $z = 0{,}50$. Na Tabela IV encontramos esta última área como sendo

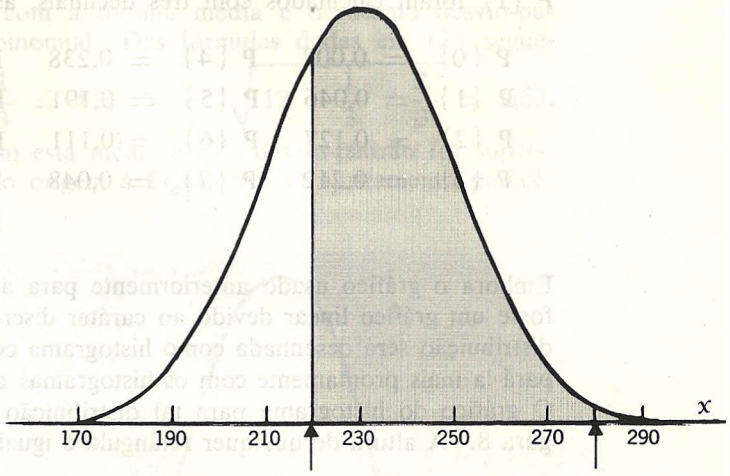


FIGURA 7. Distribuição normal particular.

0,1915; conseqüentemente, a área desejada é a soma destas duas áreas, ou seja, 0,6853. Embora as duas curvas normais das Figuras 6 e 7 pareçam muito diferentes, se usarmos a mesma escala em ambos os eixos, elas propositadamente foram desenhadas de maneira semelhante para que a equivalência das áreas correspondentes se tornasse evidente.

4 — APROXIMAÇÃO NORMAL À BINOMIAL

Os problemas relacionados com a distribuição binomial são bastante fáceis de resolver desde que o número n de tentativas não seja grande. Se n for grande, os cálculos decorrentes do uso da fórmula (1) se tornam excessivamente demorados; conseqüentemente, uma boa aproximação simples da distribuição deverá ser muito útil. Tal aproximação existe na forma da distribuição normal conveniente. Com a finalidade de investigar esta aproximação, consideremos alguns exemplos numéricos.

Seja $n = 12$ e $p = \frac{1}{3}$ e construamos o gráfico da distribuição

binomial correspondente. Pelo uso da fórmula (1), os valores de $P\{x\}$ foram calculados com três decimais, assim:

P {0} = 0,008	P {4} = 0,238	P {8} = 0,015
P {1} = 0,046	P {5} = 0,191	P {9} = 0,003
P {2} = 0,127	P {6} = 0,111	P {10} = 0,000
P {3} = 0,212	P {7} = 0,048	P {11} = 0,000
		P {12} = 0,000

Embora o gráfico usado anteriormente para a distribuição binomial fosse um gráfico linear devido ao caráter discreto da variável x, esta distribuição será desenhada como histograma com o objetivo de compará-la mais prontamente com os histogramas da distribuição normal. O gráfico do histograma para tal distribuição é apresentado na Figura 8. A altura de qualquer retângulo é igual à probabilidade dada

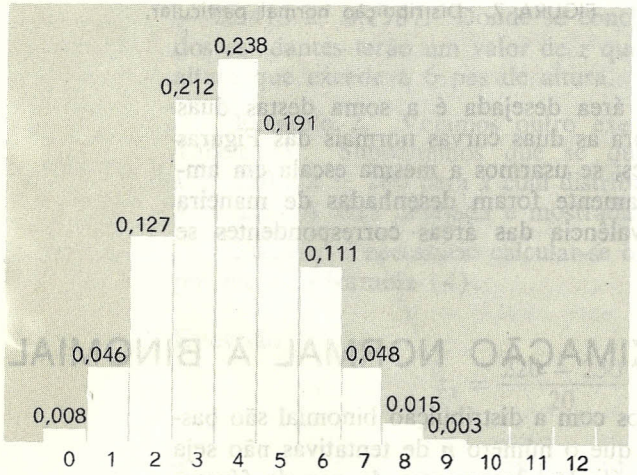


FIGURA 8. Distribuição binomial para $p = \frac{1}{3}$ e $n = 12$.

pela fórmula (5) para o ponto de classe correspondente. Já que a base do comprimento de qualquer retângulo é 1, a área de qualquer retângulo será igual a sua altura e, portanto, essas probabilidades serão também dadas pelas áreas dos retângulos correspondentes. Examinando a forma deste histograma vê-se que ele se ajusta muito bem a uma curva normal. Como uma curva de distribuição normal é completamente determinada por sua média e seu desvio-padrão, a curva normal natural que deve ser usada aqui é aquela com a mesma média e o mesmo desvio-padrão da distribuição binomial. Das fórmulas dadas em (2) segue-se que $\mu = 12 \cdot \frac{1}{3} = 4$ e $\sigma = \sqrt{12 \cdot \frac{1}{3} \cdot \frac{2}{3}} = 1{,}63$.

Uma curva normal com esta média e este desvio-padrão foi superposta à Figura 8 dando origem à Figura 9. O ajustamento parece

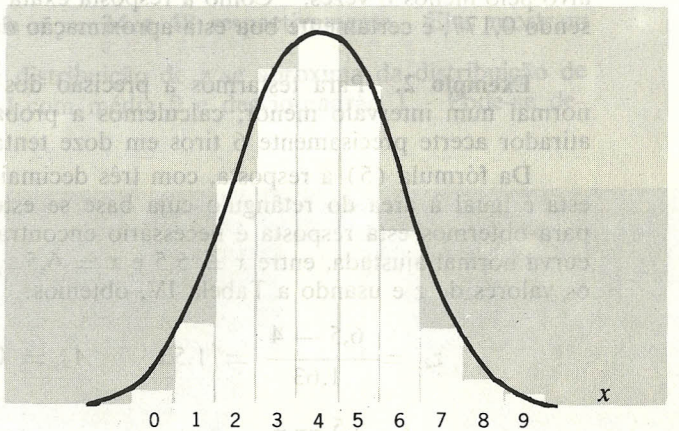


FIGURA 9. Distribuição binomial para $p = \frac{1}{3}$ e $n = 12$, com uma curva normal ajustada.

bastante bom apesar de $n = 12$ ser um valor pequeno de n e a teoria avançada só prometer bons ajustamentos para grandes valores de n.

Como teste de precisão da aproximação da curva normal e como ilustração sobre como usar os métodos da curva normal para aproximar-se das probabilidades binomiais, consideremos alguns problemas relacionados com a Figura 9.

Exemplo 1. Se a probabilidade de que um atirador atinja um alvo é de 1/3 e se ele dá 12 tiros, qual a probabilidade de que vá acertar pelo menos 6 vezes.

A resposta correta, com três decimais, será obtida pela adição dos valores, na fórmula (5) de $x = 6$ a $x = 12$, o que dará 0,177. Geometricamente, esta resposta é a área daquela parte do histograma na Figura à direita de $x = 5{,}5$. Portanto, para aproximar-se desta

probabilidade pelos métodos da curva normal, basta encontrar a área sob aquela parte da curva normal ajustada, à direita de 5,5. Uma vez que a curva normal ajustada tem $\mu = 4$ e $\sigma = 1{,}63$ segue-se que:

$$z = \frac{x - \mu}{\sigma} = \frac{5{,}5 - 4}{1{,}63} = 0{,}92.$$

Agora, da Tabela IV, a área entre $z = 0$ e $z = 0{,}92$ é de 0,321. Donde a área à direita de $z = 0{,}92$ é de 0,179; conseqüentemente esta é a aproximação desejada para a probabilidade de atingir o alvo pelo menos 6 vezes. Como a resposta exata foi calculada como sendo 0,177, é certamente boa esta aproximação da curva normal.

Exemplo 2. Para testarmos a precisão dos métodos da curva normal num intervalo menor, calculemos a probabilidade de que o atirador acerte precisamente 6 tiros em doze tentativas.

Da fórmula (5) a resposta, com três decimais, é 0,111. Como esta é igual à área do retângulo cuja base se estende de 5,5 a 6,5, para obtermos esta resposta é necessário encontrarmos a área sob a curva normal ajustada, entre $x = 5{,}5$ e $x = 6{,}5$. Assim, calculando os valores de z e usando a Tabela IV, obtemos:

$$z_2 = \frac{6{,}5 - 4}{1{,}63} = 1{,}53, \qquad A_2 = 0{,}4370$$

$$z_1 = \frac{5{,}5 - 4}{1{,}63} = 0{,}92, \qquad A_1 = 0{,}3212$$

Subtraindo estas duas áreas temos como resultado 0,116, que, comparado ao valor da probabilidade exato de 0,111, é também bom. Destes dois exemplos decorre que os métodos da curva normal dão boas aproximações mesmo para algumas situações, como as consideradas aqui, nas quais n não é muito grande.

Suponhamos, agora, que o valor de $p = \frac{1}{3}$ não seja alterado mas permite-se que n aumente em tamanho. O histograma resultante, como a da Figura 9, se deslocará para a direita, se achatará e diminuirá de altura. É difícil examinar tal histograma e ver se a curva normal, em questão, se ajusta bem a ele. Estas mudanças indesejadas no histograma podem ser evitadas voltando-se para a variável correspondente em unidades-padrão. Da fórmula (4) isto significa traçar o histograma para a variável

$$z = \frac{x - \mu}{\sigma}$$

Quando se usam as fórmulas (2), a variável-padrão z assume a forma

$$(6) \qquad z = \frac{x - np}{\sqrt{npq}}$$

Como a variável z possui uma distribuição com média 0 (zero) e desvio-padrão 1, o histograma para z permanecerá sem alteração, não se alongando e achatando com o aumento de n, como no caso do histograma para x.

As Figuras 10 e 11 apresentam os histogramas para a variável z quando $p = \frac{1}{3}$ e $n = 24$ e 48 respectivamente. Elas mostram quão rapidamente a distribuição de z se aproxima da distribuição de uma variável normal com média 0 e desvio-padrão 1. Pode-se demonstrar por métodos avançados que se p for mantido fixo e n se tornar progressivamente maior, então a distribuição de z se tornará progressivamente aproximada da distribuição de uma variável normal com média 0 e desvio-padrão 1. De um ponto de vista prático, a experiência tem demonstrado que a aproximação é muito satisfatória

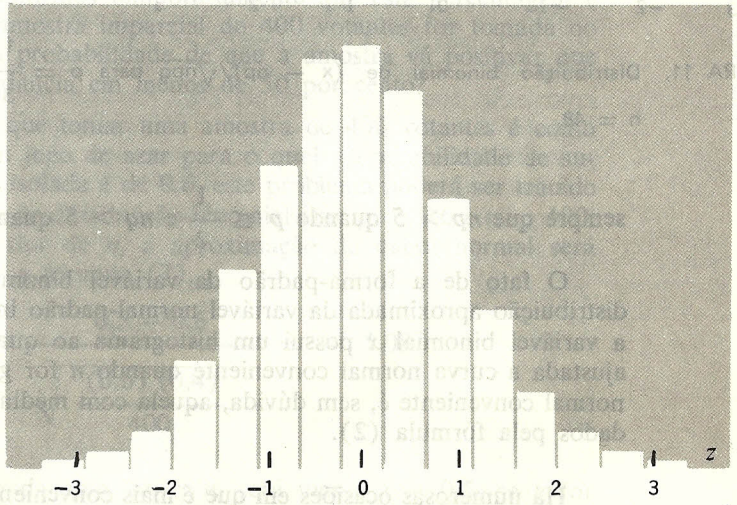


FIGURA 10. Distribuição binomial de $(x - np)/\sqrt{npq}$ para $p = \frac{1}{3}$, e $n = 24$.

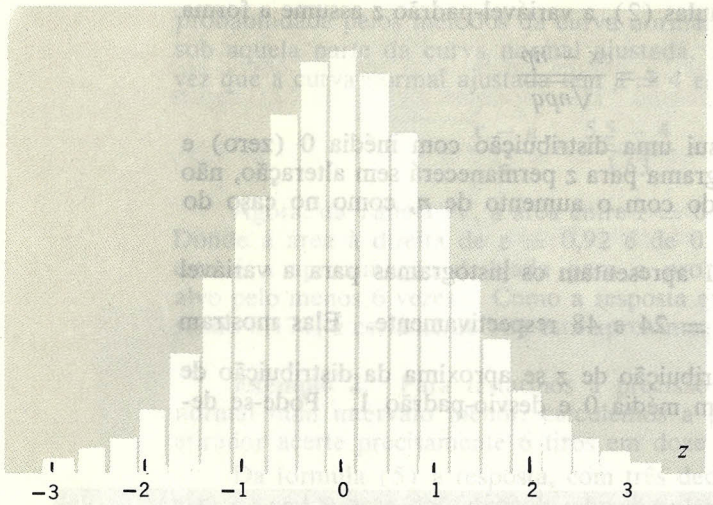


FIGURA 11. Distribuição binomial de $(x - np)/\sqrt{npq}$ para $p = \frac{1}{3}$ e $n = 48$.

sempre que $np > 5$ quando $p \leq \frac{1}{2}$ e $nq > 5$ quando $p > \frac{1}{2}$.

O fato de a forma-padrão da variável binomial possuir uma distribuição aproximada da variável normal-padrão implica então que a variável binomial x possui um histograma ao qual pode ser bem ajustada a curva normal conveniente quando n for grande. A curva normal conveniente é, sem dúvida, aquela com média e desvio-padrão dados pela fórmula (2).

Há numerosas ocasiões em que é mais conveniente trabalhar com a proporção de sucessos x/n, em n tentativas do que com o número de sucessos verificados, x. Se o numerador e o denominador na fórmula (6) forem divididos por n, z assumirá a seguinte forma

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \tag{7}$$

onde a proporção de sucessos, x/n, está simbolizada por $\hat{p}$. O valor de z não se alterou: só a forma em que foi escrito; conseqüentemente z ainda possui uma distribuição normal aproximada com média zero e desvio-padrão unitário quando n é grande. Isso implica que a proporção de sucessos, $\hat{p}$, possui um histograma que pode ajustar-se bem à curva normal quando n for grande. A curva normal conveniente é agora esta com média e desvio-padrão dados pela fórmula.

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \tag{8}$$

Exemplo 1. Como ilustração final do uso da aproximação da curva normal à distribuição binomial e do uso da fórmula (7), consideremos o problema seguinte. Suponhamos que um político diga que um levantamento no seu distrito mostrou que 60 por cento dos seus adeptos concordaram com o seu voto a respeito de uma questão legislativa. Se admitirmos temporariamente que esta percentagem é correta e se uma amostra imparcial de 400 votantes for tomada no seu distrito, qual a probabilidade de que a amostra vá positivar que houve uma concordância em menos de 50 por cento?

Se se admitir que tomar uma amostra de 400 votantes é como jogar 400 vezes um jogo de azar para o qual a probabilidade de sucesso numa jogada isolada é de 0,6, este problema poderá ser tratado como um problema de distribuição binomial com $p = 0{,}6$ e $n = 400$. Para tão grande valor de n, a aproximação da curva normal será excelente. Usando a fórmula (7),

$$z = \frac{0{,}5 - 0{,}6}{\sqrt{\frac{(0{,}6)\,(0{,}4)}{400}}} = -4{,}08$$

Agora a proporção da amostra, x/n, será menor que 0,5, se z for menor que $-4{,}08$. A probabilidade de que $z < -4{,}08$ é por simetria igual à probabilidade de que $z > 4{,}08$, a qual é considerada demasiado pequena para constar da Tabela IV. Assim, se devia acontecer que menos de 50 por cento da amostra fosse favorecer o político, sua afirmativa anterior de 60 por cento certamente seria desacreditada.

Pode-se levantar objeções, e com razão, de que obter uma amostra de 400 votantes não é equivalente a jogar um jogo de azar 400 vezes. Há questões concernentes à independência das tentativas e à constância da probabilidade a que se deve responder antes que se

possa ficar completamente satisfeito com o modelo da distribuição binomial para este problema.

Exemplo 2. Como segundo exemplo do uso da fórmula (7), vamos resolver o problema que segue. Uma experiência com exame para principiantes de inglês mostrou que só 50 por cento dos estudantes conseguiam passar. Se uma nova classe de 200 principiantes fizer o exame, qual a percentagem de que passe nele pelo menos 55 por cento?

Aqui

$$z = \frac{0{,}55 - 0{,}50}{\sqrt{\dfrac{(0{,}5)\,(0{,}5)}{200}}} = 1{,}41$$

Mas $x \geq 0{,}55$ se, e somente se, $z \geq 1{,}41$. Portanto, pela Tabela IV a probabilidade de $x \geq 0{,}55$ é de 0,08.

Exemplo 3. Como ilustração final, consideremos o seguinte problema. Suponhamos que 5 por cento dos indivíduos que foram inoculados contra um surto de gripe tiveram reações sérias indesejáveis devido à inoculação. Usando a aproximação normal, calcule a probabilidade de que mais de 8 por cento dos 200 indivíduos inoculados tenham tais reações.

Se a fórmula (7) for aplicada, calcula-se primeiro

$$z = \frac{0{,}08 - 0{,}05}{\sqrt{\dfrac{(0{,}05)\,(0{,}95)}{200}}} = \frac{0{,}03}{\sqrt{0{,}000237}} = \frac{0{,}03}{0{,}0154} = 1{,}95$$

Da Tabela IV segue-se que $P\{z > 1.95\} = 0{,}0256$, que é a probabilidade desejada de que mais de 8 por cento tenha a reação.

Não se deve presumir dos exemplos anteriores que todos os problemas de distribuição binomial possam ser tratados satisfatoriamente por meio da aproximação normal, mesmo sendo n bastante grande. Por exemplo, se $n = 80$ e $p = \frac{1}{20}$ o cálculo dos valores de $P\{x\}$ e o gráfico destes valores mostrarão que a distribuição não é suficientemente simétrica para permitir um bom ajustamento à curva normal. Como a média aqui é 4 e x não pode assumir valores negativos, um excesso de distribuição está concentrado perto do zero. Considerações como esta deram origem a uma regra empírica para uma boa aproximação, estabelecida no parágrafo (6).

5 — OBTENÇÃO DA DISTRIBUIÇÃO BINOMIAL

Para aqueles que estudaram a seção 11 do Capítulo 3, a correção da fórmula (1) pode ser demonstrada como segue.

Consideremos um experimento de n tentativas e uma seqüência particular de ocorrências que produzirá precisamente x sucessos e $n - x$ insucessos. Tal seqüência é a seguinte, na qual todos os sucessos ocorrem primeiro, seguidos dos insucessos:

$$\overbrace{\mathrm{SS}\cdots\mathrm{S}}^{x}\ \overbrace{\mathrm{FF}\cdots\mathrm{F}}^{n-x}.$$

Outra seqüência é a seguinte, na qual ocorreu primeiro um insucesso, seguido por x sucessos consecutivos depois pelos insucessos restantes.

$$\mathrm{F}\overbrace{\mathrm{SS}\cdots\mathrm{S}}^{x}\ \overbrace{\mathrm{FF}\cdots\mathrm{F}}^{n-x-1}.$$

Por causa da independência das tentativas, a probabilidade de se obter a primeira destas duas seqüências é dada por

$$\overbrace{p\cdot p\cdots p}^{x}\cdot\overbrace{q\cdot q\cdots q}^{n-x} = p^x q^{n-x}$$

A probabilidade da segunda seqüência é dada por

$$q\cdot\overbrace{p\cdot p\cdots p}^{x}\cdot\overbrace{q\cdot q\cdots q}^{n-x-1} = p^x q^{n-x}$$

Assim a probabilidade é a mesma para as duas seqüências e será a mesma para toda seqüência que satisfaça a condição de se ter x sucessos e $n - x$ insucessos.

O número de modo pelo qual o evento desejado pode ocorrer é igual ao número de diferentes seqüências que podem ser escritas, do tipo já mostrado, das que contêm x letras S e $n-x$ letras F. Mas este é igual às diversas maneiras de escolhermos x posições de n posições disponíveis nas quais colocar a letra S. As posições $n-x$ restantes terão automaticamente a letra F. As n posições podem ser numeradas e tratadas como n cartas numeradas. O problema, então, é escolher o x dessas cartas. Desde que somente o número sobre as cartas têm interesse e não a ordem em que são tiradas, trata-se de um problema de combinação. Do estudo feito na seção 11 do Capítulo 3, o número de maneiras de se escolher x cartas de n cartas distintas é dado pela fórmula da combinação (13) do Capítulo 3, a saber:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Este é portanto, o número de seqüências que produz exatamente x sucessos. Como cada uma dessas seqüências representa uma das maneiras mutuamente exclusivas em que o evento desejado pode ocorrer e cada uma de tais seqüências tem a mesma probabilidade de ocorrer, a saber $p^x\, q^{n-x}$, segue-se que a probabilidade desejada é obtida adicionando-se esta probabilidade tantas vezes quantas são as seqüências. Mas o número de tais seqüências já foi encontrado como sendo $\binom{n}{x}$; portanto $P\{x\}$ é obtido multiplicando-se $p^x\, q^{n-x}$ por $\binom{n}{x}$ que confirma a fórmula (1).

6 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Suponhamos que 60 por cento dos estudantes de uma grande universidade trabalhem em tempo parcial. Façamos x simbolizar o número desses estudantes numa amostra de 8 estudantes escolhidos ao acaso dos registros de matrícula. Supondo tão grande o tamanho da universidade que $p = 0{,}6$ possa ser considerado constante quando se tem uma amostra de apenas 8 alunos, (a) 1 — ache uma expressão para $P\{x\}$ que dê a distribuição de x (b) 1 — Use o resultado de (a) para calcular $P\{x \geq 6\}$. (c) 2 — Use as fórmulas (2) para calcular a média e o desvio-padrão de x. (d) 4 — Use a aproximação normal para aproximar o valor encontrado em (b). Eis as soluções:

(a) A amostra de 8 pode ser tratada como 8 tentativas independentes de um experimento para o qual $p = 0{,}6$; donde x é uma variável binomial com distribuição dada por

$$P\{x\} = \frac{8!}{x!\,(8-x)!}\,(0{,}6)^x(0{,}4)^{8-x}$$

(b) $P\{x \geq 6\} = P\{6\} + P\{7\} + P\{8\}$

$$= \frac{8!}{6!\,2!}\,(0{,}6)^6\,(0{,}4)^2 + \frac{8!}{7!\,1!}\,(0{,}6)^7\,(0{,}4)^1 +$$

$$+ \frac{8!}{8!\,0!}\,(0{,}6)^8\,(0{,}4)^0 =$$

$$= (0{,}6)^6\,[28(0{,}4)^2 + 8(0{,}6)\,(0{,}4) +$$

$$+ (0{,}6)^2] = 0{,}315$$

(c) $\mu = 8(0{,}6) = 4{,}8$, $\sigma = \sqrt{8(0{,}6)\,(0{,}4)} = 1{,}39$

(d) $x = \dfrac{5{,}5 - 4{,}8}{1{,}39} = 0{,}50$, $P\{z > 0{,}50\} = 0{,}31$

Exemplo 2. Um restaurante pode acomodar 50 clientes. A experiência mostra que 10 por cento dos que fazem reservas não comparecerão. Suponhamos que o restaurante aceite 55 reservas. Utilizemos x para simbolizar o número de clientes que comparecem. (a) 1 — Ache uma expressão para $P\{x\}$ que dê a distribuição de x. (b) 2 — Calcule a média e o desvio-padrão de x por meio das fórmulas (2). (c) 4 — Use a aproximação normal para calcular a probabilidade de que o restaurante possa acomodar todos os clientes que comparecerem.

(a) As 55 reservas podem ser tratadas como 55 tentativas independentes de um experimento para o qual $p = \frac{9}{10}$; donde x é uma variável binomial cuja distribuição é dada por

$$P\{x\} = \frac{55!}{x!\,(55-x)!}\,(0{,}9)^x\,(0{,}1)^{55-x}$$

(b) $\mu = 55\,(0{,}9) = 49{,}5$, $\sigma = \sqrt{55(0{,}9)\,(0{,}1)} = 2{,}2$

(c) Todos os clientes serão acomodados se no máximo 50 deles comparecerem. O problema é pois cacular $P\{x \leq 50\}$. Isto pode ser obtido achando-se a área sob o histograma de $P\{x\}$ à esquerda de 50,5. O valor correspondente de z é

$$z = \frac{50{,}5 - 49{,}5}{2{,}2} = 0{,}45$$

Donde

$$P\{x \leq 50\} = P\{z \leq 0{,}45\} = 0{,}67$$

Exemplo 3. Façamos x simbolizar o peso em variedades de peso de um salmão-rei capturado na desembocadura de certo rio e suponhamos que x possua uma distribuição normal com média 30 e desvio-padrão 6. Calcule a probabilidade de que, se um pescador pegar um salmão, seu peso (a) 3 — seja pelo menos de 41 unidades de peso, (b) 3 — esteja entre 20 e 40 unidades de peso, inclusive.

(a) $z = \dfrac{41 - 30}{6} = 1{,}83$, $P\{x \geq 41\} = P\{z \geq 1{,}83\} = 0{,}03$

(b) $z_1 = \dfrac{20 - 30}{6} = -1{,}67$, $z_2 = \dfrac{40 - 30}{6} = 1{,}67$,

$$P\{20 \leq x \leq 40\} = P\{-1{,}67 \leq z \leq 1{,}67\} = 2P\{0 \leq z \leq 1{,}67\} = 0{,}90$$

Exemplo 4. Uma máquina desenhada para perfurar placas de ferro é disposta para fazer um buraco de 1,02 polegadas de diâmetro com garantia de que o buraco não tenha menos que 1 polegada de diâmetro. (a) 3 — Se o desvio-padrão dos buracos feitos por esta máquina for de 0,01 polegadas, que percentagem de buracos será menor que 1 polegada? (b) 3 — Se placas com buracos maiores que 1,05 polegadas não forem utilizáveis, que percentagem de placas que contêm 2 buracos cada uma será rejeitada, porque pelo menos um dos buracos será grande demais ou pequenos demais?

(a) Suponhamos que os diâmetros tenham uma distribuição normal. O problema é calcular $P\{x > 1{,}00\}$, onde x é a variável normal. Aqui

$$z = \frac{1{,}00 - 1{,}02}{0{,}01} = -2{,}00$$

Donde $P\{x < 1{,}00\} = P\{z < -2{,}00\} = 0{,}023$

(b) $$z = \frac{1{,}05 - 1{,}02}{0{,}01} = 3{,}00$$

$$P\{x > 1{,}05\} = P\{z > 3{,}00\} = 0{,}001$$

Há, portanto, uma probabilidade de 0,024 de que o buraco não seja satisfatório. A probabilidade de que ambos os buracos sejam satisfatórios é de (0,976) (0,976) = 0,95. Portanto, 5 por cento dos buracos serão rejeitados.

EXERCÍCIOS

Seção 1

1 — Verifique os valores dados no texto, no Quadro 3, por meio da fórmula (1).

2 — Use o método de enumeração de ocorrências possíveis com suas probabilidades correspondentes para obter a distribuição binomial do número de caras conseguido no lançamento de uma moeda por 4 vezes. Verifique estes resultados por meio da fórmula (1).

3 — Resolva o problema 2 para o número de 1 que se obteve no lançamento de um dado, quatro vezes.

4 — Lança-se uma moeda 5 vezes. Usando a fórmula (1), calcule os valores de $\mathbf{P}\{x\}$, onde x simboliza o número de caras e trace o gráfico de $\mathbf{P}\{x\}$.

5 — Se a probabilidade de você ganhar uma mão de bridge é de $\frac{1}{4}$ e você joga 5 mãos, calcule os valores de $\mathbf{P}\{x\}$, onde x simboliza o número de vitórias, por meio da fórmula (1).

6 — Se 20 por cento das válvulas são defeituosas e se comprarmos uma caixa de 8 válvulas, qual a probabilidade de que pelo menos 6 dessas válvulas estejam boas?

7 — Suponhamos que duas marcas de aspirina sejam igualmente eficientes na redução da dor. Qual a probabilidade de que se pedirmos a 10 indivíduos em necessidade, para comprarem ambas as marcas, 8 deles pelo menos comprem a marca A?

8 — Doze pares de animais experimentais são submetidos a duas dietas diferentes. A atribuição da dieta aos membros de cada par é feita ao acaso. Após o experimento acha-se a diferença em ganho de peso entre os animais submetidos à dieta A e à dieta B. Se a diferença for positiva, o resultado será chamado de sucesso. Qual a probabilidade de que pelo menos 9 sucessos ocorram se as duas dietas não possuem diferenças reais no que diz respeito a propriedades de ganho de peso?

9 — No problema 6, quantas válvulas você precisaria comprar de modo a ter uma probabilidade de pelo menos 0,9 na obtenção de pelo menos 6 válvulas boas?

10 — Tendo-se um registro para os últimos 5 anos da porcentagem dos dias em que as ações da Bolsa estiveram em alta, pode-se aplicar a distribuição binomial ao cálculo da probabilidade de que as ações experimentem uma alta pelo menos durante 20 dias do mês próximo? Explique.

11 — Explicar por que não seria estritamente correto aplicar a distribuição binomial ao cálculo da probabilidade de que vá chover pelo menos 10 dias durante o próximo mês de janeiro, se cada dia de janeiro for tratado como uma tentativa de um acontecimento e houver um registro da porcentagem dos dias chuvosos nesse mês.

Seção 2

12 — Para a variável binomial para a qual $\mathbf{n} = 8$ e $\mathbf{p} = \frac{1}{4}$, calcule a média e o desvio-padrão e verifique os resultados por meio das fórmulas (2).

13 — Para o problema 5, calcule a média e o desvio-padrão da variável $\mathbf{x}$ e verifique os resultados por meio das fórmulas (2).

14 — Desvie a fórmula $\mu = np$ para a distribuição binomial geral $P\{x\}$ dada pela fórmula (1) escrevendo os termos em $\sum_{x=0}^{n} xP\{x\}$, fatorando depois o fator comum np e calculando $\sum_{x=0}^{n-1} Q(x)$ onde $Q(x)$ é o mesmo que $P\{x\}$ para $n-1$ tentativas e finalmente reconhecendo que esta última soma, que tem o valor 1, é o outro fator em $\Sigma\, x\, P\{x\}$. A fórmula parece óbvia mas sua derivação algébrica não o é.

15 — Consulte um texto de estatística mais avançada para observar a técnica algébrica usada na obtenção da fórmula $\sigma = \sqrt{npq}$.

Seção 3

16 — Sendo x normalmente distribuído com média 12 e desvio-padrão 2, use a Tabela IV para calcular a probabilidade de que (a) $x > 14$, (b) $x > 11$, (c) $x < 10$, (d) $x < 10{,}5$ (e) $10 < x < 13$.

17 — Admitindo-se que a estatura (x) de rapazes universitários é normalmente distribuída com média de 69 polegadas e desvio-padrão de 3 polegadas, usar a Tabela IV para calcular a probabilidade de que (a) $x < 66$ polegadas, (b) 65 polegadas $< x <$ 71 polegadas.

18 — Suponha que a sua nota num exame em unidades-padrão (z) é 0,8 e que se admite que as notas sejam normalmente distribuídas. Que porcentagem dos estudantes poder-se-ia esperar que tirasse notas maiores que você?

19 — Um professor de atletismo anuncia a seus alunos que dará notas individuais por um julgamento comparativo com o aproveitamento de toda a turma. Por exemplo, se, no salto em altura, ele der 20 por cento de notas máximas e se a experiência mostrar que a média é de 4 pés e 10 polegadas com o desvio-padrão de 4 polegadas, qual a altura mínima que um atleta deverá pular para conseguir a nota máxima?

20 — Supondo que os pontos atribuídos aos alunos de uma escola sejam normalmente distribuídos com média 115 e desvio-padrão 8, calcule a percentagem de estudantes que terão um QI (a) maior que 130, (b) mais baixo que 100, (c) entre 105 e 125, inclusive.

21 — A experiência com um exame de colocação em inglês básico indica que as notas do exame são distribuídas de maneira aproximadamente normal com média de 130 e desvio-padrão de 20. Se se exige a nota 100 para que se passe no exame, qual a percentagem de estudantes que se espera não passem no exame?

22 — A vida média de um motor elétrico é de 6 anos com um desvio-padrão de 2 anos. Se a amplitude da vida de tal motor pode ser tratada como variável normal e se o motor estiver garantido, quanto tempo deve valer a garantia para que, no máximo, 15 por cento dos motores falhem antes de expirar a garantia?

19. Média 1,50 m $\sigma = 0{,}1$ m $R = 1{,}58$ m

Seção 4

23 — Lança-se uma moeda 8 vezes. Ache a probabilidade, tanto exatamente por meio da fórmula (1) quanto aproximadamente por meio da proximação da curva normal, de se obter (a) 6 caras, (b) pelo menos, 6 caras.

24 — Se a probabilidade de você ganhar em determinado jogo é de 0,6, ache a probabilidade tanto exatamente por meio da fórmula (1) quanto aproximadamente por meio da aproximação da curva normal de se ganhar 4 ou mais de 7 jogos.

25 — Se 30 por cento dos estudantes têm visão defeituosa, qual a probabilidade de que pelo menos metade dos membros de uma classe de 18 alunos possua visão defeituosa? Use a aproximação de curva normal.

26 — Se 10 por cento de tubos de imagem de televisão se queimam antes de sua garantia expirar, (a) qual a probabilidade de que um comerciante que vendeu 100 desses tubos seja forçado a substituir pelo menos 14 deles? (b) qual a probabilidade de que ele substitua pelo menos 5 e não mais do que 14 tubos? Use a aproximação da curva normal.

27 — Se 20 por cento dos motoristas, em certa cidade, têm pelo menos um acidente por ano, qual a probabilidade de que a percentagem para 200 clientes de uma companhia de seguros, nesta cidade, vá exceder a 25 por cento, durante o ano seguinte? Use a aproximação da curva normal.

28 — Se 10 por cento dos alfinetes fabricados forem defeituosos, qual a probabilidade aproximada de que a percentagem de defeituosos numa caixa de 200 exceda a 15 por cento?

29 — Um exame de múltipla-escolha contém 20 questões. Cada questão tem 4 respostas possíveis, das quais uma apenas é a verdadeira. Se o estudante tentar adivinhar todas as questões, qual a probabilidade de que ele obtenha, pelo menos, 8 respostas certas?

30 — Suponhamos que 20 por cento de todas as declarações de imposto federal possuam deduções ilegais. Se escolhermos cada décimo declarante de imposto para examinarmos a validez de suas declarações, qual a probabilidade de que, no máximo, 5 de 200 declarantes sejam culpados de deduções ilegais?

31 — Suponhamos que $n = 20$ e $p = \frac{1}{10}$. Calcule $P\{0\}$ e na base de seu valor mostre que não se pode esperar aqui, para a distribuição inteira, um bom ajustamento à curva normal.

32 — Para o problema 31, calcule $P\{x=0\}$ por meio da aproximação da curva normal e compare com o resultado do problema 31.

33 — Dê um exemplo de distribuição binomial para a qual $n > 100$ e p não é pequeno, mas para o qual não haverá uma distribuição normal boa.

Geral

34 — Todo ano desde que se iniciou o campeonato entre dois times de futebol profissional, houve apostas de jogadores profissionais. Estas apostas aparecem no quadro abaixo:

Ano	Apostas	Sucesso do Jogo
1967	Green Bay por 15	Green Bay 35, Kansas City 10
1968	Green Bay por 17	Green Bay 33, Oakland 14
1969	Baltimore por 17	New York 16, Baltimore 7
1970	Minnesota por 11	Kansas City 23, Minnesota 7
1971	Baltimore por 3	Baltimore 16, Dallas 13
1972	Dallas por 5	Dallas 24, Miami 3
1973	Washington por 2	Miami 14, Washington 7
1974	Pittsburgh por 3	Pittsburgh 26, Minnesota 16

Se um apostador tivesse de utilizar os números dos apostadores profissionais e apostar com um amigo nesta base, escolhendo sempre o vencedor previsto, teria ganho 5 vezes e perdido 3. Se tivesse apostado somente no sucesso do jogo, sem especificar ponto teria ganho e perdido o mesmo número de vezes que ao apostar por pontos.

Supondo que os profissionais estejam corretos em suas previsões, o que implica a probabilidade de metade deles estar correta em cada previsão, calcule a probabilidade de que um apostador ganhe pelo menos 5 das 8 vezes se seguir as previsões.

35 — Uma caixa contém as 9 cartas seguintes: o três, o quatro e o cinco de espadas, o três e o quatro de paus, o três e o quatro de copas e o quatro e o cinco de ouro. Se tivermos que retirar uma carta da caixa e tivermos que repetir o experimento 10 vezes com reposição das cartas retiradas logo após cada lance e se **x** significar o número de cartas pretas que serão obtidas, (a) ache uma expressão para $\mathbf{P}\{x\}$ que dê a distribuição de **x**; (b) use o resultado de (a) para calcular $\mathbf{P}\{4 \leq x \leq 5\}$; (c) use a aproximação normal para aproximar o valor encontrado em (b); (d) calcule a probabilidade aproximada, usando a aproximação normal, de que a proporção de cartas pretas em 100 tentativas do experimento será menor do que 0,6.

36 — Pacotes de alimentos para o almoço são tabelados como contendo 12 onças de cereais. A máquina que enche os pratos está sujeita a erros com um desvio-padrão de 0,1 por onça.

(a) Se a máquina tiver que encher um pacote com 12,1 onças de cereal, que percentagem de pacotes terão pelo menos 0,1 onça a menos no peso?

(b) Se o produtor desejar que, no máximo, 5 por cento de seus pacotes tenham uma falta de pelo menos 0,1 onça, que peso médio deveria usar?

37 — Um avião pode acomodar 300 passageiros, 30 dos quais na primeira classe e 270 na classe de turismo. Se a companhia de aviação reservar 30 lugares de primeira classe e 290 de turismo e se a probabilidade de não comparecimento de quem faz uma reserva for de 0,1, qual a probabilidade de que todos os passageiros que comparecerem sejam acomodados, se os lugares da primeira classe puderem ser usados pelos passageiros de turismo?

6

AMOSTRAGEM

1 — INTRODUÇÃO

Os capítulos anteriores estavam relacionados com a descrição das distribuições de freqüência obtidas a partir da amostra de uma população e com a construção de distribuições de probabilidade que possam servir para representar a população.

Supôs-se que a distribuição da população poderia ser representada com grande precisão se a amostra fosse suficientemente grande e, portanto, a distribuição da freqüência amostral representasse a população de maneira satisfatória. Por exemplo, se um fabricante de camisas escolhe uma amostra grande de tamanho de colarinhos para adultos, ele presume que a distribuição desses tamanhos de sua amostra estará muito próxima da população de adultos do sexo masculino. Se isto não fosse verdade, a amostra não serviria como base satisfatória para fazermos inferências estatísticas sobre a população amostrada.

O problema de como extrair amostras, de modo a podermos fazer inferências válidas concernentes à população, já foi brevemente discutido no Capítulo 2. Ficou dito, então, que um método de amostragem chamado *amostragem aleatória* deste ponto de vista era satisfatório. Discutiremos as propriedades da amostragem aleatória que permite fornecer inferências válidas e descreveremos métodos eficientes de obtermos amostras aleatórias. Após estas considerações preliminares estudaremos o problema da determinação do grau de precisão com referência ao tamanho da amostra. Tal estudo levará em conta tanto resultados teóricos como vários experimentos amostrais. O propósito dos experimentos é demonstrar a variabilidade amostral e tornar mais plausíveis os resultados teóricos.

2 — AMOSTRAS ALEATÓRIAS

No Capítulo 2, a amostragem aleatória foi definida como um processo amostral onde cada membro da população tem a mesma chance de ser escolhido. Em termos de probabilidade isto implica que a probabilidade de qualquer membro particular que está sendo selecionado é $1/N$, onde N simboliza o número de indivíduos da população. Mais genericamente, se a amostra deve conter r indivíduos, quando se diz então que a amostra é do tamanho de r, a amostragem é definida como sendo aleatória se cada combinação de r indivíduos da população tem a mesma chance de ser selecionada.

Embora a amostragem aleatória fosse defendida antes com apoio no fato de que é um meio de amostragem cuja distribuição representa a distribuição da população em miniatura (isto é, cuja distribuição é a representação miniatural da distribuição de toda a população), há ainda razões mais importantes para defendê-la. A mais importante delas é que a amostragem aleatória leva a modelos de probabilidade para as distribuições. Como as conclusões a serem tiradas acerca das populações por meio de amostras devem ser baseadas em probabilidades, as amostras devem ser selecionadas de tal maneira que as regras de probabilidade possam ser aplicadas a elas. Para vermos como isto acontece, voltemos ao problema discutido no capítulo precedente relacionado a uma amostra de 400 votantes.

Presumiu-se então que a amostra de 400 votantes poderia ser tratada como 400 tentativas independentes de um experimento para o qual a probabilidade de sucesso de uma única tentativa era de $p = 0{,}6$. Esta suposição permitiu que o problema fosse tratado como um problema de distribuição binomial e permitiu, portanto, o cálculo de probabilidades das várias ocorrências possíveis por meio da distribuição binomial e de sua aproximação normal.

Os dois requisitos necessários que permitem que a distribuição binomial seja aplicada a problemas práticos são a independência das tentativas e a constância da probabilidade para todas as tentativas. Para uma população finita de N indivíduos, a probabilidade de que um indivíduo selecionado ao acaso favoreça a certo candidato não permanecerá necessariamente constante à medida que amostras sucessivas forem tomadas; contudo, se N for muito grande comparado ao tamanho da amostra, a proporção de indivíduos que favorecem o candidato mudará muito pouco com o progresso da amostragem.

Portanto, se a amostra for escolhida de modo que em cada etapa da amostragem cada membro restante da população tenha a mesma chance de ser escolhido, estarão cumpridas satisfatoriamente as condições para um modelo binomial. Assim, a amostragem aleatória permitirá o uso de um modelo de distribuição binomial para tal problema, contanto que N seja grande. Se N não for suficientemente grande para justificar tal uso, pode-se empregar um modelo de probabilidade mais sofisticado. Não o discutiremos, porém, aqui.

Se se empregou algum outro tipo de amostragem para se obter a amostra de 400 votantes, não haverá segurança de que um modelo de distribuição binomial seja válido para os cálculos dessa probabilidade. Por exemplo, se um pesquisador começa numa rua residencial e escolhe as primeiras 400 casas desta rua, ele poderia muito bem encontrar um grupo de indivíduos cuja semelhança de status econômico tenderia a fazer a porção deles favorecer ou desfavorecer muito o candidato, o que não representaria a oposição de toda a cidade. Em geral, não se pode fazer afirmações de probabilidade válidas sobre ocorrências de outros tipos de métodos amostrais que não os aleatórios. É por isso que os estatísticos insistem em dizer que as amostras devem ser aleatórias.

Há muitos exemplos, na vida pública, de fracassos que ocorreram porque as conclusões se apoiaram em amostras não aleatórias. Por exemplo, em 1936 o *Literary Digest* fez um inquérito pelo correio para predizer quem seria eleito, então, presidente. Enviou 10 000 000 (dez milhões) de questionários e baseado em mais de 2 000 000 de respostas predisse que London seria eleito. O que sucedeu, na realidade, é que Roosevelt recebeu aproximadamente 60 por cento dos votos naquela eleição. Os jornais regularmente informam as opiniões e predições de políticos como se suas afirmações possuíssem muito mérito, mesmo quando o político declara que as suas opiniões se baseiam na sondagem de seus eleitores. Os homens de negócio geralmente tomam decisões incorretas em campos como mercado quando suas decisões são baseadas em informações falhas obtidas de amostras pobres.

Pode-se obter amostras aleatórias empregando-se técnicas de jogo de azar. Por exemplo, se houver 10 000 votantes numa cidade e tivermos de colher uma amostra de tamanho 400, bastaria escrever o nome de cada votante numa tira de papel, misturar essas tiras num grande recipiente, e tirar depois 400 tiras do recipiente, tendo o cuidado de misturar as tiras após cada retirada.

Um método mais barato e menos trabalhoso de obtermos amostras aleatórias seria usar uma tabela de números aleatórios. Tal tabela poderia ser construída escrevendo-se os dígitos de 0 a 9 em dez tiras de papel, misturar bem as tiras entre cada retirada e registrar os dígitos obtidos. A seqüência de dígitos aleatórios resultantes poderia, então, ser usada para se construir números aleatórios de, digamos, dois dígitos cada um. O quadro de números aleatórios dado na Tabela XI, do apêndice, poderia ter sido obtido deste modo, mas foi obtido por um método mais refinado. Suponhamos agora que os 10 000 votantes estejam arrolados num registro de votantes e que cada votante esteja associado a um número de 0 000 a 9 999. Para obtermos então uma amostra de 400 nomes basta escolhermos 400 conjuntos de números de 4 dígitos na Tabela XI. Como estes números ocorrem aos pares, no quadro, basta selecionar dois pares vizinhos para conseguirmos um nome. Há 50 linhas de números em cada coluna; dessa forma, seria suficiente selecionar 16 colunas. Da maneira pela qual são formados os números aleatórios, segue-se que cada um dos números de quatro dígitos tem a mesma probabilidade de ser formado em qualquer lugar especificado da tabela; por conseqüência, dá no mesmo escolher os números sistematicamente pela leitura segundo uma coluna ou escolhê-los salteados na tabela. Poderia acontecer, é claro, que um ou mais indivíduos da população fossem selecionados mais de uma vez quando os números aleatórios fossem usados, porque esses números se formaram independentemente, e, como conseqüência, um dado número de 4 dígitos pode ocorrer várias vezes na tabela. Pode-se, então, escolher alguns números aleatórios extras para completar a amostra. Se houvesse apenas 7 500 votantes registrados no distrito, rejeitar-se-ia qualquer número de quatro dígitos maior que 7 499 obtido da tabela, porque não haveria votantes associados a esses números.

3 — ESTIMADORES NÃO TENDENCIOSOS

A finalidade de tirarmos amostras de populações é, naturalmente, a de conseguirmos informações sobre a distribuição da população. Esta informação é geralmente expressa na forma de uma distribuição de freqüência empírica com seu histograma, ou na forma de quantidades aritméticas tais como $\bar{x}$ e s^2. Assim, uma distribuição empírica é uma estimativa amostral da distribuição de probabilidades que representa a população. Do mesmo modo, $\bar{x}$ e s^2 são estimativas amostrais dos parâmetros μ e σ^2 da distribuição de probabilidades da população.

Como a maior parte das inferências estatísticas que serão feitas sobre populações por meio de amostras se relacionarão a quantidades tais como média e variância, é tempo de começarmos a estudar o

grau de segurança da média amostral $\bar{x}$ e a variância amostral s^2 como estimativas dos valores dos parâmetros populacionais correspondentes μ e σ^2.

Este capítulo tratará principalmente do estudo das propriedades da média amostral porque é a mais importante das duas e é muito mais fácil de tratar que a variância amostral.

O problema de determinarmos a precisão da estimativa amostral do parâmetro de uma população pode ser dividido em duas partes. Uma, refere-se a saber se a estimativa é tendenciosa. Por exemplo, se obtivéssemos uma estimativa da altura média dos adultos do sexo masculino de uma amostra aleatória tirada de universitários, o valor de $\bar{x}$ tenderia a ser maior do que μ, porque universitários tendem a ser maiores que o resto da população adulta mais velha. Uma estimativa que tende a superestimar ou subestimar o valor de um parâmetro é chamada *estimativa tendenciosa.* Logo teremos uma definição mais precisa.

A segunda parte do problema refere-se à determinação da distância maior ou menor da estimativa, em termos de probabilidade, do parâmetro que ele está estimando. Esta parte do problema será considerada na próxima seção.

Com o propósito de discutirmos a propriedade de uma estimativa não tendenciosa, consideremos o experimento segundo o qual tomamos uma amostra aleatória de tamanho 10 de alguma população cuja distribuição de probabilidade tenha a média μ. Façamos $\bar{x}$ significar a média da amostra. Se este experimento fosse repetido várias vezes, disporíamos de um grande número de valores de $\bar{x}$ para obtermos uma distribuição empírica da variável $\bar{x}$. Esperar-se-ia que a média dessa distribuição empírica se aproximasse de algum valor fixo com a repetição indefinida do experimento. Se o valor, ao se aproximar, novamente fosse μ, diríamos então que $\bar{x}$ era uma estimativa não tendenciosa de μ. Em outras palavras, $\bar{x}$ é uma estimativa não tendenciosa de μ se a média de sua distribuição de probabilidade for μ. Como o valor esperado E de uma variável aleatória é a média de sua distribuição de probabilidade, segue-se que $\bar{x}$ será uma *estimativa não tendenciosa* de μ se, e somente se, seu valor esperado for igual a μ. Esta definição é válida para outros parâmetros e para seus estimadores correspondentes.

Ora, pode-se demonstrar usando as três propriedades de E dadas no Capítulo 4 que

$$E[\bar{x}] = \mu \qquad \text{e} \qquad E[s^2] = \sigma^2$$

Assim, $\bar{x}$ é uma estimativa não tendenciosa de μ, e s^2 uma estimativa não tendenciosa de σ^2. Esta último fato justifica-se por que dividimos $\sum_{i=1}^{n} (x_i - \bar{x})^2 f_i$ por n-1 e não por n na definição de s^2 dada pela fórmula (7), Capítulo 2. Se n tivesse sido usado no denominador, s^2 teria sido uma estimativa tendenciosa de σ^2 e teria produzido uma estimativa menor que σ^2, na média.

Como ilustração, suponhamos que se escolha um grande número de amostras de tamanho 5 e que se calcule $\sum_{i=1}^{5} (x_i - \bar{x})^2/5$ para cada amostra. O valor médio desses estimadores se aproximará de $\frac{4}{5}\sigma^2$, ao passo que se usássemos 4 como denominados o valor médio ter-se-ia aproximado de σ^2.

4 — DISTRIBUIÇÃO DE $\bar{x}$ COMO AMOSTRA DE UMA POPULAÇÃO NORMAL

Em vista do fato de $\bar{x}$ e s^2 serem estimativas não tendenciosas de μ e σ^2, o estudo de sua precisão se reduz ao estudo de como estas estimativas variam em torno de μ e σ^2, respectivamente, sob experimentos amostrais repetidos. Como $\bar{x}$ é mais útil que s^2 em problemas de inferência estatística e como uma distribuição amostral é mais fácil de estudar que a de s^2, somente estudaremos a distribuição de x, neste capítulo.

Há dois teoremas matemáticos básicos sobre a precisão de $\bar{x}$. Ao invés de expô-los simplesmente, efetuaremos dois experimentos amostrais para observar como $\bar{x}$ varia em amostragem repetida e verificar que os resultados amostrais concordam com os que constam dos teoremas.

No primeiro experimento, tomaremos amostras de tamanho 4 de uma população cuja distribuição de probabilidade é dada pela Figura 1. Esta distribuição é uma aproximação discreta à distribuição normal padrão. Foi obtida usando-se a Tabela IV para calcular a percentagem da área sob a curva normal padrão em intervalos unitários com o intervalo do meio centrado na origem. Incidentalmente, as percentagens obtidas da Figura 1 (quando as duas percentagens dos intervalos extremos são combinadas com os intervalos vizinhos de cada

lado) dão as percentagens que são usadas freqüentemente por instrutores que "as classificam sobre a curva" para determinar a percentagem das notas A, B, C, D e F.

Tomaram-se amostras de tamanho 4, desta distribuição discreta, por meio de números aleatórios encontrados na Tabela XI. Primeiro construiu-se uma forma de tabulação do tipo mostrado na Tabela I. Por este processo todos os pares de dígitos aleatórios foram divididos em sete grupos conforme as proporções mostradas na Figura 1 e associados aos pontos de classe da Figura 1.

Por exemplo, $x = -3$ é atribuído ao par 00, que é 1 por cento de tais pares e o valor $x = -2$ é atribuído aos pares de 01

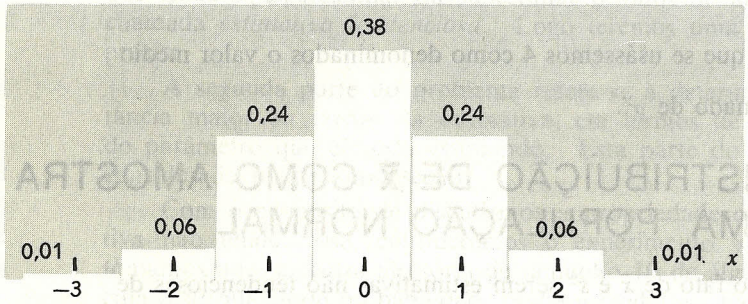


FIGURA 1. Aproximação à distribuição normal padrão.

QUADRO 1

x	Números Aleatórios	1	2	3	...
−3	00				
−2	01-06		/	/	
−1	07-30	/			
0	31-68	//	/	///	
1	69-92	/	/		
2	93-98		/		
3	99				
x		0	$\frac{1}{4}$	$-\frac{1}{2}$	...

a 06 inclusive, que inclui 6 por cento de todos os pares de números aleatórios. Quatro desses pares de números aleatórios são lidos na Tabela XI e registrados no intervalo de classe conveniente para experimento.

Os resultados dos três primeiros desses experimentos aparecem no Quadro 1. Este experimento foi repetido 100 vezes. Depois, a média de tal experimento foi calculada e registrada na última linha do Quadro 1, onde se vê $\bar{x}$.

Os valores de $\bar{x}$ foram tabulados para fornecerem as freqüências mostradas no Quadro 2 no qual a terceira linha dá as percentagens, em fração decimal, das freqüências absolutas correspondentes.

QUADRO 2

$\bar{x}$	$-\frac{5}{4}$	$-\frac{4}{4}$	$-\frac{3}{4}$	$-\frac{2}{4}$	$-\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{4}{4}$	$\frac{5}{4}$
f	1	4	8	11	15	19	18	12	5	4	3
$\frac{f}{100}$	0,01	0,04	0,08	0,11	0,15	0,19	0,18	0,12	0,05	0,04	0,03

Finalmente, esta tabela de freqüências foi expressa como um histograma. Como as áreas totais devem ser iguais a 1 ao compararmos distribuições de probabilidade, e como o intervalo de classe desta tabela é de 1/4, é necessário desenhar retângulos que sejam quatro vezes maiores que os valores de f/100. O histograma resultante, com área 1, aparece na Figura 2.

Comparando as Figuras 1 e 2, é evidente que as médias amostrais baseadas em quatro medições cada uma não variam tanto quanto os valores amostrais individuais. Isto era de se esperar porque, por exemplo, um grande valor de $\bar{x}$ requereria quatro valores grandes de x e a probabilidade de se obter quatro valores grandes é muito menor que a probabilidade de se obter um único valor grande. O desvio-padrão da distribuição de $\bar{x}$ é sem dúvida muito menor que o da distribuição de x. Além disso, é evidente que a distribuição de $\bar{x}$ possui uma média próxima de 0, que é a média da distribuição de x. Finalmente, parece que a distribuição de $\bar{x}$, exceto para a diferença em dispersão, possui uma distribuição do mesmo tipo aproximado normal da distribuição de x.

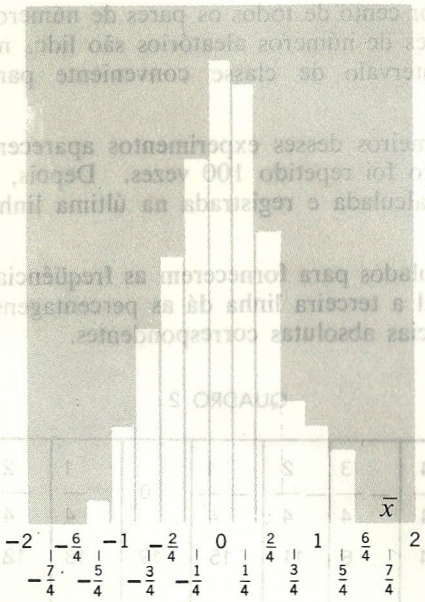


FIGURA 2. Distribuição de $\bar{x}$ para amostras de tamanho 4 de distribuição da Figura 1.

Examinando a Figura 1 vemos claramente que $\mu = 0$. Cálculos baseados na fórmula (2) do Capítulo 4 e as probabilidades dadas pela Figura 1 mostram que $\sigma = 1{,}07$. Se a distribuição da Figura 1 tivesse sido a de uma variável normal padrão contínua em vez de uma aproximação discreta a ela, o valor de σ teria sido exatamente 1.

Os cálculos da média e do desvio-padrão para a distribuição de $\bar{x}$ empírica dada pelo Quadro 2 foram feitos por meio das fórmulas (4) e (8) do Capítulo 2 e produziram os valores de 0,015 e 0,545, respectivamente. Assim, a média da distribuição $\bar{x}$ está muito próxima da média, $\mu = 0$, da distribuição de x, ao passo que o desvio-padrão da distribuição de $\bar{x}$ é aproximadamente a metade do desvio-padrão, $\sigma = 1{,}07$, da distribuição x.

Se este experimento tivesse sido feito, digamos, 500 vezes e não apenas 100, as irregularidades como as da Figura 2 desapareceriam e descobrir-se-ia que as propriedades da distribuição $\bar{x}$ que acabamos de discutir seriam muito mais evidentes. Descobriríamos, assim, que o histograma se ajustaria muito bem a uma curva normal, a média da distribuição de $\bar{x}$ estaria muito mais próxima de 0 e o desvio-padrão da distribuição de $\bar{x}$ teria um valor muito próximo da metade do valor do desvio-padrão da distribuição de x. Embora as amostras aqui tenham sido tiradas de uma distribuição normal aproximada como a da Figura 1 e não de uma distribuição normal exata, poderíamos obter resultados semelhantes se a aproximação se tornasse sempre melhor pela escolha de um intervalo de classe pequeno, ao invés de um intervalo de comprimento 1 como o da Figura 1.

Felizmente não é necessário efetuarmos tais experimentos amostrais repetidos para chegarmos a uma distribuição teórica de $\bar{x}$. Usando regras de probabilidade e métodos matemáticos avançados, é possível obter a equação da curva representativa da distribuição de $\bar{x}$ quando a amostragem for tirada de uma distribuição normal exata e não de uma aproximação. Isto corresponde ao que foi feito no Capítulo 4 para chegarmos a uma distribuição teórica para a binomial x sem realizarmos experimentos amostrais. Acontece que $\bar{x}$ possuirá então uma distribuição normal se x a possuir, com a mesma média de x mas com um desvio-padrão que é $1/\sqrt{n}$ vezes o desvio-padrão de x. Estes resultados matemáticos são expressos em forma de um teorema.

(1) **Teorema 1.** *Se* x *possui uma distribuição normal com média* μ *e desvio-padrão* σ, *então a média amostral* $\bar{x}$, *baseada numa amostra aleatória de tamanho* n, *possuirá também uma distribuição normal com média* μ *e desvio-padrão* $\sigma/\sqrt{n}$.

A distribuição de $\bar{x}$ dada por este teorema é muitas vezes chamada distribuição amostral de $\bar{x}$ por causa de sua relação com experimentos amostrais repetidos, mesmo no caso de ser obtida por métodos puramente matemáticos.

Os resultados do experimento amostral que acabamos de estudar parecem concordar plenamente com este teorema. O histograma da Figura 2 parece o tipo de histograma que se obtém de amostras de uma população normal e como $\sigma \doteq 1$ e $n = 4$ aqui, sua média e desvio-padrão estão em concordância com os valores teóricos de 0 e $\sigma/\sqrt{4} \doteq \frac{1}{2}$ dados pelo teorema. Desse teorema pode-se tirar a conclusão de que as médias de amostras de tamanho 4, de uma população normal, possuem somente a metade da variabilidade em torno da média da população que as medidas individuais possuem.

Exemplo. Para vermos como este teorema pode ser usado para determinar a precisão de $\bar{x}$ como estimativa de μ quando x possui distribuição normal, consideremos o seguinte problema.

Façamos x representar a altura de um indivíduo escolhido aleatoriamente de uma população de adultos do sexo masculino. Suponhamos que x possua distribuição normal com média $\mu = 68$ polegadas e desvio-padrão $\sigma = 3$ polegadas. A geometria desta distribuição aparece na Figura 3. Dados sobre estatura de pessoas do sexo

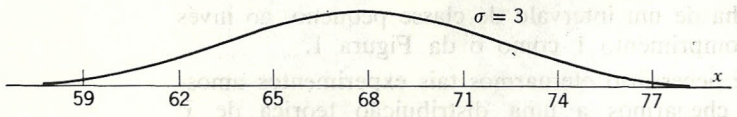


FIGURA 3. Distribuição normal de x.

masculino mostram que estas suposições são bastante realistas. O problema a ser resolvido é o seguinte: se tomarmos uma amostra de tamanho $n = 25$ a partir de uma população, qual a probabilidade de que a média amostral $\bar{x}$ difira da média da população em menos de uma polegada?

Como $\mu = 68$, $\sigma = 3$ e $n = 25$, segue-se do teorema 1 que $\bar{x}$ possuíra uma distribuição normal com média 68 e desvio-padrão dado por

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0{,}6$$

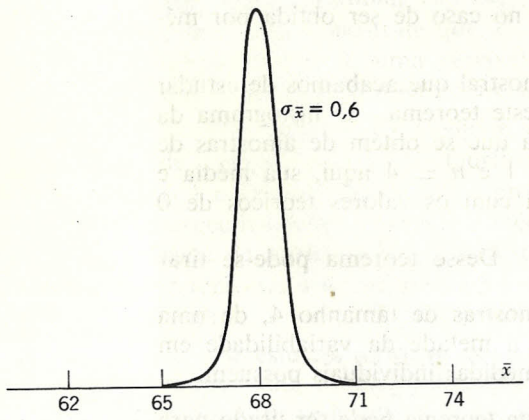


FIGURA 4. Distribuição normal de $\bar{x}$ quando n= 25.

A geometria desta distribuição aparece na Figura 4. Ora o erro da estimativa será menor que uma polegada se $\bar{x}$ cair dentro do intervalo (67,69). Os valores de z correspondentes a $\bar{x} = 67$ e $\bar{x} = 69$ são dados por

$$z_1 = \frac{67 - 68}{0{,}6} = 1{,}67; \text{ e } z_2 = \frac{69 - 68}{0{,}6} = 1{,}67$$

Como, da Tabela IV, a probabilidade de que z caia entre $-1{,}67$ e 1,67 é igual a 0,90, este valor é também a probabilidade de que x esteja em erro, por um máximo de 1 polegada. Assim, escolhendo uma amostra aleatória de tamanho 25 e usando o valor resultante de $\bar{x}$ para estimar a média da população total, podemos estar certos de que esta estimativa não diferirá da média da população em mais de uma polegada.

5 — DISTRIBUIÇÃO DE $\bar{x}$ COMO AMOSTRA DE UMA POPULAÇÃO NÃO NORMAL

Suponhamos agora que a variável x não possua uma distribuição normal. O que poderá ser dito sobre a distribuição de $\bar{x}$? Muitos estatísticos efetuaram experimentos amostrais com diferentes espécies de distribuições não normais de x para ver que efeito esta não-normalidade teria sobre a distribuição de $\bar{x}$. O resultado surpreendente foi sempre que, se n for maior que cerca de 25, a distribuição de $\bar{x}$ aparecerá normal apesar da distribuição da população escolhida para x. Esta propriedade notável de $\bar{x}$ é de grande importância prática, porque grande número de problemas reais envolve amostras suficientemente grandes para permitir que se suponha que $\bar{x}$ esteja normalmente distribuído, e possibilitar assim o uso dos métodos da conhecida curva normal para resolver problemas relacionados com médias sem estarem ligados à natureza da distribuição da população.

Um conhecido teorema matemático estabelece essencialmente que sob um reduzido número de suposições a distribuição de $\bar{x}$ aproximar-se-á de uma distribuição normal com o aumento do tamanho n da amostra. Este teorema pode ser expresso da seguinte maneira.

(2) **Teorema 2.** *Se* x *possui uma distribuição com média* μ *e desvio-padrão* σ, *então a média amostral* $\bar{x}$, *baseada numa amostra aleatória de tamanho* n, *terá uma distribuição que se aproxima da distribuição de uma variável normal com média* μ *e desvio-padrão* $\sigma/\sqrt{n}$ *quando* n *tende ao infinito.*

Para mostrarmos que $\bar{x}$ possuirá uma distribuição normal aproximada com média μ e desvio-padrão $\sigma/\sqrt{n}$, mesmo no caso de x não estar normalmente distribuída, se n for suficientemente grande, foi levado a efeito o experimento amostral que segue. Foram escolhidas amostras aleatórias de tamanho 10 da população representada pelo histograma da Figura 5. A variável aqui é uma variável dis-

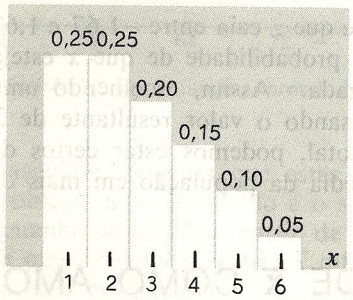


FIGURA 5. Distribuição da população para experimento amostral.

creta que pode assumir somente valores inteiros de 1 a 6, com as probabilidades correspondentes indicadas no histograma. O Teorema 2 não requer que x seja uma variável contínua. Pode ser contínua ou discreta. Usa-se um histograma em vez de um gráfico linear apenas a fim de melhor expor a falta de normalidade.

Números aleatórios de dois dígitos da Tabela XI foram divididos em 6 grupos correspondentes aos 6 possíveis valores de x. Aos primeiros 25 por cento desses números, a saber, de 00 a 24, foi atribuído o valor 1 de x. Aos seguintes 25 por cento, isto é, de 25 a 49, foi atribuído o valor 2 de x, aos seguintes 20 por cento, ou seja, de 50 a 69, o valor 3 de x etc...

Após fazer essas atribuições de valores, foi selecionada uma coluna de números aleatórios de dois dígitos. Os primeiros 10 números dessa coluna deram lugar à primeira amostra de tamanho $n = 10$ da população de x. O segundo conjunto de 10 números dessa coluna deu lugar à segunda amostra de 10 etc...

Isso foi continuando, com tantas colunas quantas foram necessárias, até serem extraídas as 100 amostras.

Na fase seguinte do experimento foram calculados os valores de $\bar{x}$ para aquelas 100 amostras. Depois que os valores de $\bar{x}$ foram

calculados, foram classificados numa tabela de freqüência como foi discutido no Capítulo 2. Os resultados dessa classificação aparecem no Quadro 3 e na Figura 6. Não se fez esforço para manter a

QUADRO 3

f	1	0	2	7	13	16	19	16	14	9	1	2
$\bar{x}$	1,55	1,75	1,95	2,15	2,35	2,55	2,75	2,95	3,15	3,35	3,55	3,75

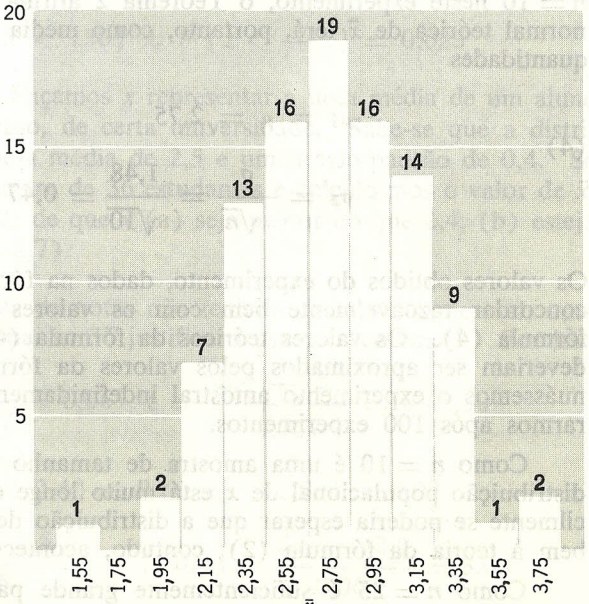


FIGURA 6. Histograma de 100 médias amostrais.

área do histograma igual a 1, porque só interessa, no caso, a forma do histograma. O histograma da Figura 6 tem, sem dúvida, a aparência de um histograma obtido a partir da amostragem de uma população normal. Assim a afirmação do Teorema (2) de que $\bar{x}$ deveria ter uma distribuição normal aproximada parece ter sido cumprida aqui.

Finalmente, os valores da média e o desvio-padrão para os dados do Quadro 3 foram calculados por métodos explicados no Capítulo 2, com os seguintes resultados:

(3) média da distribuição $\bar{x} = 2{,}77$
desvio-padrão da distribuição $\bar{x} = 0{,}41$

Estes valores precisam ser comparados com os valores esperados do Teorema 2. Para fazermos esta comparação, é necessário conhecermos os valores de μ e de σ para a população x que está sendo amostrada. Cálculos por meio das fórmulas (1) e (2) do Capítulo 4 para a distribuição de probabilidades apresentada pela Figura 5 deram os valores $\mu = 2{,}75$ *e* $\sigma = 1{,}48$, com duas decimais. Como $n = 10$ neste experimento, o Teorema 2 afirma que a distribuição normal teórica de $\bar{x}$ terá, portanto, como média e desvio-padrão as quantidades

(4)
$$\mu_{\bar{x}} = \mu = 2{,}75$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1{,}48}{\sqrt{10}} = 0{,}47$$

Os valores obtidos do experimento, dados na fórmula (3), parecem concordar razoavelmente bem com os valores teóricos dados na fórmula (4). Os valores teóricos da fórmula (4) são aqueles que deveriam ser aproximados pelos valores da fórmula (3) se continuássemos o experimento amostral indefinidamente ao invés de pararmos após 100 experimentos.

Como $n = 10$ é uma amostra de tamanho pequeno e como a distribuição populacional de x está muito longe de ser normal, dificilmente se poderia esperar que a distribuição de $\bar{x}$ se ajustasse tão bem à teoria da fórmula (2); contudo, aconteceu isso claramente.

Como $n = 25$ é suficientemente grande para justificar o uso do Teorema 2 no problema usado para ilustrar o Teorema 1, segue-se que não teria sido necessário supor que as estaturas dos adultos do sexo masculino fossem normalmente distribuídas para se resolver esse problema.

Os teoremas anteriores serão usados extensivamente nos capítulos seguintes para fazermos inferências estatísticas sobre médias populacionais.

6 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Suponhamos que os pontos, x, num teste de colocação sejam normalmente distribuídos com média 160 e desvio-padrão 20. Se tirarmos uma amostra de tamanho 16 e calcularmos o valor de $\bar{x}$, qual é a probabilidade de que (a) 4-$\bar{x}$ exceda 165 (b) 4-$\bar{x}$ seja menor que 150?

(a) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{16}} = 5,\ z = \frac{165 - 160}{5} = 1$$
$$P\{\bar{x} > 165\} = P\{z > 1\} = 0{,}16$$

(b) $$z = \frac{150 - 160}{5} = -2$$
$$P\{\bar{x} < 150\} = P\{z < -2\} = 0{,}023$$

Exemplo 2. Façamos x representar a nota média de um aluno selecionado, ao acaso, de certa universidade. Sabe-se que a distribuição de x tem uma média de 2,5 e um desvio-padrão de 0,4. Se tomarmos uma amostra de 36 estudantes e calcularmos o valor de $\bar{x}$, qual a probabilidade de que $\bar{x}$ (a) seja menor do que 2,4, (b) esteja no intervalo (2,4, 2,7).

(a) Embora médias de notas não tenham distribuição normal, uma amostra de tamanho 36 justifica o uso do Teorema 2. Por isso

$$\sigma_{\bar{x}} = \frac{0{,}4}{\sqrt{36}} = 0{,}067,\ z = \frac{2{,}4 - 2{,}5}{0{,}067} = -1{,}50$$
$$P\{\bar{x} < 2{,}4\} = P\{z < -1{,}50\} = 0{,}07$$

(b) $$z_1 = \frac{2{,}4 - 2{,}5}{0{,}067} = -1{,}50,\ z_2 = \frac{2{,}7 - 2{,}5}{0{,}067} = 3{,}00$$
$$P\{2{,}4 < \bar{x} < 2{,}7\} = P\{-1{,}50 < z < 3{,}00\} =$$
$$= P\{0 < z < 1{,}50\} + P\{0 < z < 3{,}00 =$$
$$= 0{,}4332 + 0{,}4987 = 0{,}93$$

EXERCÍCIOS

1 — Sugerir como tomar uma amostra aleatória de 100 estudantes de uma universidade.

2 — Dê razões pelas quais tomar cada décimo nome da letra A de uma lista telefônica poderia ou não ser considerado um esquema satisfatório de amostragem aleatória a fim de se estudar a distribuição da renda dos adultos numa cidade.

3 — As linhas aéreas freqüentemente deixam questionários nos bolsos das poltronas dos seus aviões para obter informações de seus clientes com respeito aos seus serviços. Criticar esse método de obter informações.

4 — Durante um prolongado debate sobre um projeto de lei importante no Senado dos Estados Unidos, o senador A recebeu 300 cartas elogiando-o pela sua posição e 100 cartas censurando-o pela mesma. O senador A considerou estas cartas como uma indicação clara do sentimento público em relação ao projeto. Comente este fato.

5 — Uma firma de negócios enviou questionários para uma amostra aleatória de 1 000 donas-de-casa de certa cidade a fim de saber suas opiniões sobre guardanapos de papel. Destas, 400 responderam. Seriam essas 400 respostas satisfatórias para julgar as opiniões gerais das donas-de-casa sobre guardanapos?

6 — Como você poderia usar números aleatórios para tomar amostras de trigo de um trigal, se este trigal tem a forma de um quadrado, cada lado medindo 1 000 pés e se cada amostra é tomada escolhendo-se um ponto aleatório no quadrado e colhendo-se o trigo dentro de um arco com 5 pés de diâmetro cujo centro é o ponto aleatório?

7 — Uma agência deseja tomar uma amostra de 200 adultos de certa seção residencial de uma cidade. Dispõe-se a fazê-lo tomando uma amostra aleatória de 200 famílias de uma lista de todas as famílias do distrito e então selecionando, ao acaso, 1 adulto de cada família. Este modo de proceder levará ou não a amostras aleatórias? Por quê?

8 — Sugerir como você poderia estabelecer um esquema aproximado de amostras aleatórias para colher amostras de (a) árvores numa floresta; (b) batatas de um caminhão de transporte carregado de sacas de batata; (c) crianças de uma comunidade abaixo dos 5 anos, que tenham tido sarampo. Em cada caso indicar alguma variável que poderia ser estudada.

Seção 4

9 — Sendo **x** normalmente distribuída com média 20 e desvio-padrão 4, calcular a probabilidade de que a média amostral $\bar{x}$, baseada numa amostra cujo tamanho é 64, (a) excederá 21; (b) excederá 19,5 (c) cairá entre 9 e 21 (d) excederá 22.

10 — Sendo **x** normalmente distribuída com média 25 e desvio-padrão 8, calcular a probabilidade de que a média amostral $\bar{x}$, baseada numa amostra de tamanho 16, (a) seja menor que 26; (b) exceda 31; (c) exceda 24; (d) seja menor que 21; (e) caia entre 28 e 29.

11 — Trace numa mesma folha de papel o gráfico de uma curva normal com média 10 e desvio-padrão 2 e o gráfico da curva média correspondente para uma amostra de tamanho 9.

12 — Como seria o gráfico da curva $\bar{x}$ do problema 11 se o tamanho da amostra fosse 36?

Seção 5

13 — Se o desvio-padrão dos pesos das crianças do primeiro grau é de 7 libras, qual é a probabilidade de que o peso médio de uma amostra aleatória de 100 dessas crianças difira em mais de 1 libra do peso médio de todas as crianças?

14 — Uma nova dieta para aumento de peso deve ser aplicada a uma amostra aleatória de 25 frangos de um lote. Se se espera que seja de 2 onças o desvio-padrão do ganho em peso no período de um mês, qual a probabilidade de que a média desta amostra difira da média do lote total, se eles são tratados pela nova dieta e não são melhores que os antigos em mais do que 1/2 onça?

15 — O peso médio dos calouros de certa universidade, nos últimos 5 anos, é de 154 libras e o desvio-padrão é de 20 libras. Se o peso médio dos primeiros 100 estudantes a serem matriculados é de 159 libras, você teria razão para crer que a nova classe de calouros é mais pesada que a normal? Dê algumas explicações possíveis.

16 — Mostrar que aproximadamente 50 por cento dos valores de uma população normalmente distribuída, cuja média é 0, cai no intervalo de $-0{,}6745\sigma$ a $0{,}6745\sigma$. O desvio $0{,}6745\sigma$ chama-se **desvio-provável.**

17 — Verificar os valores de $\bar{x}$ e **s** dada na fórmula (3) do texto.

18 — Verificar os valores de μ e σ usados na fórmula (4) do texto.

19 — Explicar como o Teorema 2 justifica o pressuposto feito no Capítulo 5 para efeito de que a distribuição normal é uma boa aproximação para a distribuição binomial se **n** for suficientemente grande. Faça isto considerando a distribuição binomial para uma amostra de tamanho **um,** caso em que $x = 0$ ou 1 com probabilidades **q** e **p,** respectivamente, e usando a fórmula (7) do Capítulo 5 com $\bar{x}/n = x$. Aqui $\mu = p$ e $\sigma = \sqrt{pq}$.

Geral

20 — Realize um experimento amostral do tipo estudado para que o Teorema 2 pareça plausível. Tome 50 amostras de tamanho 5 da distribuição discreta dada por

x	0	1	2
$P\{x\}$	0,4	0,2	0,4

Trace o histograma da distribuição $\bar{x}$ e calcule sua média e seu desvio-padrão. Calcule a média e o desvio-padrão da distribuição **x** e compare seus resultados de $\bar{x}$ com os esperados do Teorema 2. Como **n** é muito pequeno aqui e a distribuição de **x** está longe de ser normal, você não devia esperar que a distribuição de $\bar{x}$ parecesse por demais semelhante à distribuição normal.

21 — Faça cada membro da classe realizar o seguinte experimento 10 vezes. Selecionar da Tabela XI do apêndice 10 números aleatórios de um dígito e calcular sua média. Trazer estas 10 médias experimentais para classe, onde o conjunto total das médias experimentais pode ser classificado, o histograma desenhado e a média e o desvio-padrão calculados. Estes resultados deviam então ser comparados com a teoria, da mesma maneira que o experimento do texto. A distribuição populacional aqui tem $\mu = 4{,}5$ e $\sigma = 2{,}87$.

22 — Os dados da Tabela XII do apêndice podem ser usados para realizarmos experimentos amostrais e observarmos que tipos de distribuições essas variáveis possuem.

7

ESTIMATIVA

1 — ESTIMATIVAS POR PONTO E POR INTERVALO

A introdução do Capítulo 1 estabeleceu que um dos problemas fundamentais da estatística é a estimativa das propriedades das populações. Agora que as distribuições de freqüências das populações já foram estudadas, pelo menos numa extensão limitada, é possível discutir as propriedades das populações que podem ser estimadas. As duas distribuições de freqüências de população estudadas foram a distribuição binomial e a normal; conseqüentemente consideremos suas propriedades.

A distribuição binomial dada pela fórmula (1), Capítulo 5, é completamente determinada pelo número n de tentativas, e a probabilidade p de sucesso numa tentativa isolada. Os símbolos n e p são chamados *parâmetros* da distribuição. Os valores atribuídos aos parâmetros determinam a distribuição binomial particular desejada. Desde que os parâmetros n e p determinem completamente a distribuição, qualquer propriedade de uma distribuição binomial fica também completamente determinada por eles. Além disso, uma vez que o número de tentativas n é quase sempre escolhido antecipadamente nos problemas de estimativa, esses problemas de estimativa para as distribuições binomiais podem usualmente ser reduzidos ao problema de estimar p.

A distribuição normal dada pela Figura 3, Capítulo 5, é completamente determinada pelos dois parâmetros μ e σ. Os problemas de estimativas para populações normais podem, por conseguinte, normalmente ser reduzidos aos problemas de estimar μ e σ.

Há dois tipos de estimativas de parâmetros de uso comum na estatística. Um é chamado estimativa por ponto e o outro, estimativa por intervalo. Uma estimativa por ponto é a espécie familiar de estimativa, isto é, é um número obtido de computações sobre os valores da amostra que serve como uma aproximação do parâmetro estimado. Por exemplo, a proporção da amostra x/n de eleitores que favorece certo candidato é uma estimativa por ponto da proporção p da população. Semelhantemente, a média da amostra $\bar{x}$ é uma estimativa por ponto da média da população μ. Uma estimativa por intervalo para um parâmetro é um intervalo determinado por dois números obtidos de computações sobre os valores da amostra que se espera contenha o valor do parâmetro, no seu interior. A estimativa por intervalo é usualmente construída de tal maneira que a probabilidade de o intervalo conter o parâmetro possa ser especificada. A vantagem da estimativa por intervalo é que ela mostra quão precisamente o parâmetro está sendo estimado. Se o comprimento do intervalo é muito pequeno, uma alta precisão foi atingida. Tais estimativas por intervalo são chamadas intervalos de confiança. Neste capítulo serão determinadas ambas as estimativas, por ponto e por intervalo, para os parâmetros das distribuições binomial e normal.

2 — ESTIMATIVA DE μ

Consideremos o seguinte problema. Um fabricante de tabletes de vitamina C deseja comprovar a qualidade de seu produto. Ele descobriu por experiência que o conteúdo dos tabletes de vitamina C, para certo lote fabricado, tem uma distribuição aproximadamente normal. Descobriu também que o conteúdo médio para os lotes varia de um para outro, mas o desvio-padrão $\sigma = 20$, permanece quase constante, independentemente do valor da média. Se o conteúdo médio de um lote for muito baixo, o fabricante não poderá vender o lote; portanto, é importante que ele tenha uma estimativa precisa da média do lote. Para estimar o conteúdo médio de um novo lote ele testa uma amostra aleatória de 25 tabletes tomados da linha de produção e descobre que seu conteúdo médio de vitamina C é de $\bar{x} = 260$ unidades. Com estes dados e a informação anterior ao alcance, serão resolvidos três tipos de problemas de estimação.

Problema 1. Qual é o grau de precisão de $\bar{x} = 260$ como uma estimativa por ponto da média do lote μ?

Para se resolver este problema, faz-se uso da teoria apresentada no capítulo precedente. Da teoria apresentada na fórmula (2), Capítulo 6, segue-se que a média da amostra $\bar{x}$ pode ser considerada como tendo distribuição normal com média μ e desvio-padrão dado por

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = 4$$

Freqüentemente chamamos este desvio-padrão erro-padrão da média. Um esboço desta distribuição aparece na Figura 1. Como há uma probabilidade de 0,95 de que uma variável normal assuma algum valor dentro dos dois desvios-padrão de sua média (mais precisamente 1,96 desvios-padrão, com duas decimais pela Tabela IV do apêndice), segue-se que a probabilidade de que $\bar{x}$ vá assumir algum

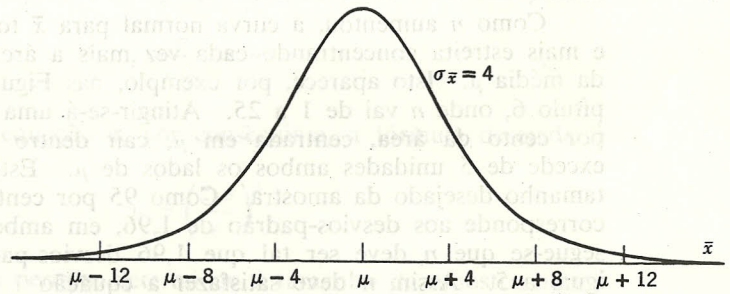


FIGURA 1. Distribuição de $\bar{x}$ para o conteúdo de vitamina C dos tabletes.

valor dentro de 8 unidades de μ é de 0,95. Como $\bar{x} = 260$ é o valor observado aqui, o fabricante pode sentir-se muito seguro de que este valor difira do valor μ da população em menos de 8 unidades, porque, a longo prazo, só em 5 por cento de tais experimentos amostrais o valor $\bar{x}$ diferirá de μ em mais de 8 unidades. A grandeza da diferença $\bar{x}$-μ é chamada erro de estimativa. Nesta linguagem, pode-se dizer que há uma probabilidade de 0,95 de que o erro de estimativa seja menor do que 8 unidades. Se fossem desejados níveis mais altos de probabilidade, poder-se-ia usar, digamos, um intervalo de três desvios-padrão para ambos os lotes de μ e então estabelecer que a probabilidade de que o erro de estimativa seja menor que 12 unidades é de 0,997.

Como $\bar{x}$ está baseado numa amostra não é possível estabelecer qual o grau de proximidade de $\bar{x}$ com relação à média populacional μ quando μ é desconhecido. Só é possível falarmos em linguagem de probabilidade, estabelecendo-se o grau provável de proximidade de $\bar{x}$

com referência a μ. Assim, o erro exato de estimativa, a saber, $\bar{x}-\mu$ só é conhecido quando μ é conhecido. Como o problema de estimativa só surge quando μ é desconhecido, deve-se introduzir probabilidades para se discutir a grandeza de um erro de estimativa, em problemas estatísticos. Mesmo se o verdadeiro valor de μ não for conhecido, pode-se ainda falar de quão próximo $\bar{x}$ está de μ e, portanto, pode-se dizer algo sobre a magnitude do erro de estimativa, $\bar{x}-\mu$, contanto que tal afirmação seja vazada em linguagem probabilística adequada.

Problema 2. Suponhamos que o fabricante não esteja satisfeito com a precisão de sua estimativa baseada na amostra de 25. Qual o tamanho da amostra adicional que ele deveria tomar de modo que pudesse estar razoavelmente certo, digamos, com uma probabilidade de 0,95 de que sua estimativa não estava errada em mais de 5 unidades?

Como n aumentou, a curva normal para $\bar{x}$ tornou-se mais alta e mais estreita concentrando cada vez mais a área nas vizinhanças da média μ. Isto aparece, por exemplo, nas Figuras 3 e 4 do Capítulo 6, onde n vai de 1 a 25. Atingir-se-á uma etapa quando 95 por cento da área, centrada em μ, cair dentro do intervalo que excede de 5 unidades ambos os lados de μ. Este valor de n é o tamanho desejado da amostra. Como 95 por cento da área central corresponde aos desvios-padrão de 1,96, em ambos os lados de μ, segue-se que n deve ser tal que 1,96 desvios-padrão para $\bar{x}$ seja igual a 5. Assim n deve satisfazer a equação

$$1{,}96\sigma_{\bar{x}} = 5$$

Como $\sigma = 20$ aqui, esta equação é equivalente a

$$1{,}96\frac{20}{\sqrt{n}} = 5$$

Resolvendo a equação para n teremos que $n = 61{,}5$. O fabricante deverá, portanto, tomar uma amostra adicional de 37, visto que ele já tomou 25, para conseguir a desejada precisão de estimativa.

Como 1,96 é um número difícil de manipular em equações desse tipo, podemos arredondá-lo para 2. A solução da equação torna-se então $n = 64$. Ainda que esta aproximação forneça uma diferença de 2, na resposta, não devemos preocupar-nos com ela, numa amostra total de 64. Além disso, o objetivo aqui é ensinar métodos estatísticos e qualquer esforço para poupar energia em cálculos trabalhosos e aplicá-la em reflexão valerá a pena.

Se alguém desejar que seja alguma outra, que não 0,95, a probabilidade de que o erro de estimativa não exceda 5, então será necessário substituir o fator 2 (ou 1,96) na equação precedente pelo

conveniente valor de z encontrado na Tabela IV, que corresponde à probabilidade desejada. Dessa forma, se fosse selecionada a probabilidade 0,90, o valor de z da Tabela IV seria 1,64. Isso decorre do fato de 90 por cento da área de uma curva normal estar compreendida dentro de 1,64 desvios-padrão a partir da média. A equação a ser resolvida para n tornar-se-ia então

$$1{,}64\frac{20}{\sqrt{n}} = 5$$

A solução desta equação é $n = 43$.

Visto que é aborrecido ter sempre de resolver uma equação desse tipo, pode-se obter outra fórmula que fornece n de maneira mais direta. Se o erro máximo permissível da estimativa for indicado por e e o valor de z correspondente à probabilidade desejada for indicado por z_0, então a equação que deve ser resolvida para n é dada por:

$$z_0\frac{\sigma}{\sqrt{n}} = e$$

A solução dessa equação e, por conseguinte, a fórmula desejada é dada por:

$$n = \left(\frac{z_0\sigma}{e}\right)^2$$

Esta fórmula permite determinar o tamanho da amostra necessária a fim de estimar μ com um grau de precisão desejado, antes de ser tomada uma única amostra, desde que seja conhecido o valor de σ. Não é necessário dispor-se de uma amostra preliminar, como no problema resolvido quatro parágrafos atrás. Se, contudo, não se conhece σ de outras fontes nem se possui uma boa estimativa dele, então é necessário tomar uma amostra preliminar de modo a permitir uma estimativa sua que possa ser usada na fórmula da determinação do tamanho de n.

Problema 3. Consideremos um terceiro tipo de problema de estimativa para este mesmo exemplo. Qual é o intervalo de confiança de 95 por cento, para μ com base na amostra original de 25?

Se for admitido que x tenha uma distribuição exatamente normal, fica claro pela teoria apresentada na fórmula (1) no Capítulo 6, ou Figura 1, que se pode escrever

$$(1) \qquad P\{\mu - 8 < \bar{x} < \mu + 8\} = 0{,}95$$

Esta é uma proposição algébrica de probabilidade daquilo que foi estabelecido em linguagem geométrica, no problema 1, a saber,

que há uma probabilidade de 0,95 de que o ponto sobre o eixo $\bar{x}$ da Figura 1 correspondente à média da amostra $\bar{x}$ não se afastará mais do que 8 unidades do ponto que representa a média μ da população. Agora é possível inventarmos esta geometria e esta álgebra e afirmarmos que há uma probabilidade de 95% de que o ponto μ não se afaste mais de 8 unidades do ponto correspondente à média $\bar{x}$ da amostra. A proporção aqui é relativa; se um ponto está dentro das 8 unidades do segundo ponto, então o segundo ponto estará dentro das 8 unidades do primeiro ponto. Essa inversão algébrica de papéis dos dois pontos será realizada agora algebricamente pelo uso das propriedades das inequações.

Uma inequação como essa $\bar{x} < \mu + 8$ pode ser reagrupada da mesma maneira que uma igualdade, só que se multiplicarmos ambos os termos da inequação por um número negativo inverteremos o sinal da inequação. Assim, a inequação $2 < 5$ torna-se a inequação $-2 > -5$ quando multiplicamos a primeira por -1. A inequação $\bar{x} < \mu + 8$ será equivalente à inequação $\bar{x} - 8 < \mu$ somando-se -8 a ambos os termos da inequação. Do mesmo modo $\mu - 8 < \bar{x}$ é equivalente a $\mu < \bar{x} + 8$. Se combinarmos esses dois resultados, veremos que a inequação dupla

$$\mu - 8 < \bar{x} < \mu + 8$$

será equivalente a essa outra inequação dupla

$$\bar{x} - 8 < u < \bar{x} + 8$$

Como conseqüência, a proposição de probabilidade (1) é equivalente à proposição de probabilidade

$$(2) \qquad P\{\bar{x} - 8 < \mu < \bar{x} + 8\} = 0{,}95$$

Em palavras, dizemos que há 0,95 de probabilidade de que a média μ populacional esteja contida dentro do intervalo que se entende de $\bar{x} - 8$ a $\bar{x} + 8$. Este intervalo é escrito da seguinte maneira: $(\bar{x} - 8,\ \bar{x} + 8)$.

Embora as fórmulas (1) e (2) sejam proposições de probabilidade equivalentes, possuem interpretações um pouco diferentes, em termos de freqüências relativas, em repetições sucessivas desse experimento de amostragem. Para cada um de tais experimentos de amostragem obtém-se um valor de $\bar{x}$. Se estes valores de $\bar{x}$ forem representados como pontos, como é visto na Figura 2, então a interpretação de freqüência da fórmula (1) é que, em tais experiências de amostragem repetidas, 95 por cento dos pontos cairão dentro da faixa representada na Figura 2.

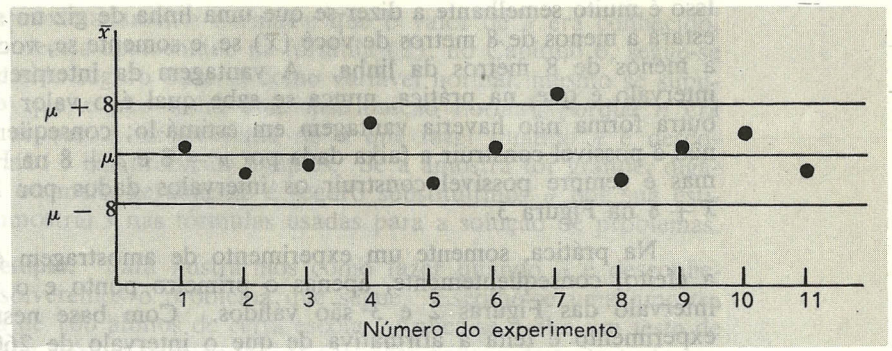


FIGURA 2. Experimentos amostrais repetidos para $\bar{x}$.

Uma interpretação de freqüência para a fórmula (2) requer que o intervalo estendendo-se de $\bar{x} - 8$ a $\bar{x} + 8$ correspondente a cada experimento da amostragem seja representado graficamente. Isto foi feito na Figura 3 para as experiências que levaram à Figura 2.

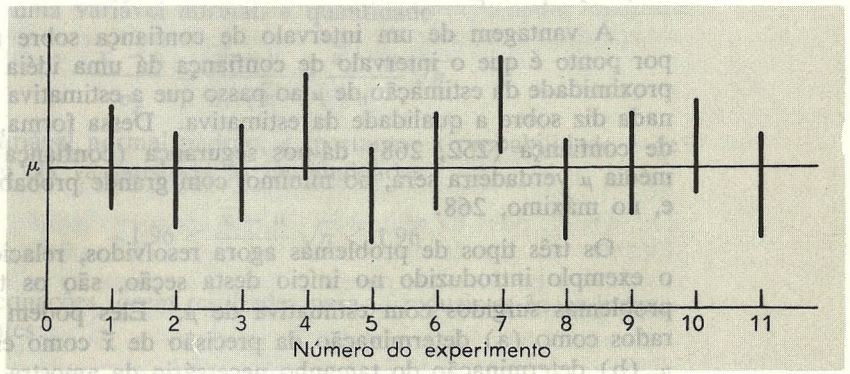


FIGURA 3. Intervalos para experimentos amostrais repetidos.

A interpretação da freqüência da fórmula (2), então, é que — em tais experiências repetidas de amostragem — 95 por cento dos intervalos conterão μ. Geometricamente está claro nas Figuras 2 e 3 que um intervalo da Figura 3 conterá μ se, e somente se, o ponto correspondente da Figura 2 recair dentro da faixa delimitada ali.

Isso é muito semelhante a dizer-se que uma linha de giz no solo (μ) estará a menos de 8 metros de você ($\bar{x}$) se, e somente se, você estiver a menos de 8 metros da linha. A vantagem da interpretação do intervalo é que, na prática, nunca se sabe qual é o valor de μ; de outra forma não haveria vantagem em estimá-lo; conseqüentemente não é possível construir a faixa dada por $\mu - 8$ e $\mu + 8$ na Figura 2, mas é sempre possível construir os intervalos dados por $\bar{x} - 8$ e $\bar{x} + 8$ na Figura 3.

Na prática, somente um experimento de amostragem é levado a efeito; conseqüentemente, apenas o primeiro ponto e o primeiro intervalo das Figuras 2 e 3 são válidos. Com base neste único experimento é feita a afirmativa de que o intervalo de $260 - 8$ a $260 + 8$ ou de 252 a 268 contém a média μ da população. Usando-se os símbolos de desigualdade, podemos escrever isto assim:

$$252 < \mu < 268$$

Se para cada um destes experimentos for feita a mesma estimativa para o intervalo correspondente, então 95 por cento de tais afirmativas serão verdadeiras na longa seqüência desses experimentos. Em vista desta propriedade, o intervalo de 252 a 268 é chamado um *intervalo de confiança* de 95 por cento para μ. Os pontos extremos do intervalo, a saber, 252 e 268 são chamados *limites de confiança* para μ.

A vantagem de um intervalo de confiança sobre a estimativa por ponto é que o intervalo de confiança dá uma idéia do grau de proximidade da estimação de μ ao passo que a estimativa por ponto $\bar{x}$ nada diz sobre a qualidade da estimativa. Dessa forma, o intervalo de confiança (252, 268) dá-nos segurança (confiança) de que a média μ verdadeira será, no mínimo, com grande probabilidade, 252 e, no máximo, 268.

Os três tipos de problemas agora resolvidos, relacionados com o exemplo introduzido no início desta seção, são os três maiores problemas surgidos com estimativa de μ. Eles podem ser enumerados como (a) determinação da precisão de $\bar{x}$ como estimativa de μ, (b) determinação do tamanho necessário da amostra para atingir determinada precisão para a estimativa de μ, e (c) determinação de um intervalo de confiança para μ.

3 — APROXIMAÇÕES

Os métodos usados para resolver os problemas anteriores eram muito simples porque se supôs que a variável x tinha uma distribuição aproximadamente normal e que o valor de σ era conhecido. Ora, pela teoria apresentada na fórmula (2), no Capítulo 6, segue-se que teria sido seguro tratar $\bar{x}$ como variável normal, mesmo se supuséssemos que x não tinha uma distribuição normal, porque n era grande, no caso. O problema de o que fazer quando σ é desconhecido não é um problema simples. Se a amostra for grande, digamos, 25 ou mais, geralmente é seguro substituirmos σ por sua estimativa amostral s nas fórmulas usadas para a solução de problemas.

Exemplo. Para ilustrarmos como fazer quando σ é desconhecido, resolveremos o problema que segue. Escolhe-se uma amostra aleatória de 100 alunos de certa escola. Aplica-se a eles um teste de inteligência para determinação do QI. Os pontos deste teste forneceram os valores amostrais $\bar{x} = 112$ e $s = 11$. Qual é um intervalo de confiança de 95 por cento para o quociente de inteligência médio na escola, baseando-se nesses valores da amostra? O problema será resolvido da mesma maneira anterior, só que será usado s em lugar de σ.

Como anteriormente, segue-se da Tabela IV que a probabilidade é de 0,95 de que a variável normal-padrão z vá satisfazer as desigualdades

$$-1{,}96 < z < 1{,}96$$

Mas se $\bar{x}$ é uma variável normal, a quantidade

$$\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma}\sqrt{n}$$

será uma variável normal padrão, e, portanto, a probabilidade é de 0,95 de que ela vá satisfazer as desigualdades

$$-1{,}96 < \frac{\bar{x} - \mu}{\sigma}\sqrt{n} < 1{,}96$$

Se essas inequações forem resolvidas para μ, reduzir-se-ão às inequações seguintes:

$$(3) \qquad \bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Este resultado pode ser usado como fórmula para obter intervalos de confiança de 95 por cento para médias de população.

Para o problema em consideração, $n = 100$, $\bar{x} = 112$ e $s = 11$. Como o valor de μ ou σ não é conhecido aqui, deve ser aproximado por sua estimativa da amostra $s = 11$. Se esses valores fossem substituídos em (3), eles assumiram a forma:

$$112 - 1{,}96\frac{11}{\sqrt{100}} < \mu < 112 + 1{,}96\frac{11}{\sqrt{100}}$$

Essas quantidades se reduzem a 109,8 e 114,2; dessa forma o desejado intervalo de confiança aproximado de 95 por cento para μ é dado por

$$109{,}8 < \mu < 114{,}2$$

Isso é apenas um intervalo de confiança aproximado de 95 por cento, pois σ foi substituído por sua aproximação da amostra s, e x não foi admitido como possuindo distribuição normal. Para uma amostra de tamanho 100, os erros surgidos em virtude dessas aproximações poderão ser desprezados.

Se alguém desejasse, digamos, um intervalo de confiança de 90 por cento em lugar de um de 95 por cento, seria preciso apenas substituir o número 1,96 pelo número 1,64 nas fórmulas precedentes, exatamente como no problema da determinação de n. Para o problema anterior os limites se tornariam então

$$112 - 1{,}64\frac{11}{\sqrt{100}} < \mu < 112 + 1{,}64\frac{11}{\sqrt{100}}$$

Se essas inequações forem simplificadas, o desejado intervalo de confiança aproximado de 90 por cento para μ tornar-se-á

$$110{,}2 < \mu < 113{,}8$$

Qualquer outra porcentagem para intervalo de confiança pode ser obtida de maneira semelhante, por meio da Tabela IV.

Os métodos de estimativa explanados nesta seção são chamados métodos de grande amostra sempre que σ for substituído por sua estimativa da amostra, porque eles são estritamente válidos somente para grande amostra. Se a amostra for menor que cerca de 25 e o valor de σ não for conhecido, esses métodos são de precisão duvidosa e, por conseqüência, necessita-se de um método mais refinado. Um método destinado a resolver tais problemas de pequenas amostras é apresentado na seção 5.

4 — ESTIMATIVA DE p

A seção 2 estava relacionada à estimativa do parâmetro para distribuições de variáveis contínuas. Esta seção explica como resolver tipos semelhantes de problemas para o parâmetro p associado às distribuições binomiais. Os métodos apresentados aqui são métodos de grande amostra, porque requerem a substituição de p por sua estimativa da amostra e também porque pressupõem que a aproximação da curva normal para a distribuição binomial seja satisfatória.

Como exemplo para ilustrar os vários tipos de problemas de estimativa a serem resolvidos, consideremos o problema de estimar a percentagem de adultos do sexo masculino em certa cidade, que fumem pelo menos um maço de cigarros por dia. Suponhamos que uma amostra aleatória de tamanho 300 fornecesse 36 de tais indivíduos. Usando estes dados, os três problemas seguintes podem ser resolvidos: (a) Qual a precisão da proporção da amostra como uma estimativa de p? (b) Que tamanho deve ter a amostra para que a probabilidade seja de 0,95 de que o erro da estimativa não vá exceder a 0,02 unidades? (c) Qual é um intervalo de confiança de 95 por cento para p?

Todos esses problemas serão resolvidos da mesma maneira que na seção 2 porque o tamanho dessa amostra é suficientemente grande para justificar o uso dos métodos da curva normal.

Problema 1. Da fórmula (7), Capítulo 5, segue-se que a proporção da amostra x/n, que será indicada por $\hat{p}$, pode ser considerada como tendo distribuição aproximadamente normal com média p o desvio-padrão

$$\sqrt{\frac{pq}{n}} = \sqrt{\frac{pq}{300}}$$

Como resultado, a probabilidade é aproximadamente 0,95 de que $\hat{p}$ vá recair dentro de 1,96 de tais desvios-padrão de p. Dessa forma, a probabilidade é aproximadamente 0,95 de que o erro da estimativa venha a ser menor que

$$(4) \qquad 1{,}96\sqrt{\frac{pq}{300}}$$

Uma vez que p é desconhecido, deve ser estimado por

$$\hat{p} = \frac{x}{n} = \frac{36}{300} = 0{,}12$$

O valor da fórmula (4) assume então o valor aproximado de

$$1{,}96\sqrt{\frac{(0{,}12)\ (0{,}88)}{300}} = 0{,}037$$

Pode-se por conseguinte estabelecer que a probabilidade é de aproximadamente 0,95 de que a estimativa da amostra $\hat{p}$ não vá diferir de p em mais de 0,037 unidades. Esse resultado dá uma boa idéia da precisão do valor de 0,12 da amostra como estimativa de p.

Problema 2. Para resolver o problema de qual o tamanho da amostra necessário para atingir-se determinada precisão para a estimativa de p, emprega-se o mesmo raciocínio usado para μ na seção 2. Isto significa que n deve ser escolhido de forma que o número conveniente de desvios-padrão de $\hat{p}$ vá igualar o erro máximo da estimativa desejado. Como anteriormente, seja e o erro máximo da estimativa escolhido e z_0 o valor de z correspondente à probabilidade desejada de não exceder este erro máximo. Então n deve satisfazer a equação

$$z_0\sqrt{\frac{pq}{n}} = e$$

A resolução desta equação para n dará a fórmula

$$n = \frac{z_0^2 pq}{e^2} \tag{5}$$

Para o problema particular considerado aqui, $e = 0{,}02$ e $z_0 = 1{,}96$. Como p é desconhecido, deverá ser estimado pelo valor da amostra $\hat{p} = 0{,}12$. Se estes valores forem substituídos na fórmula (5) o valor de n será encontrado como aproximadamente 1014; dessa forma, será necessária uma amostra adicional de aproximadamente 714 para obter-se a desejada precisão da estimativa.

Problema 3. Para se encontrar um intervalo de confiança para p, usa-se o mesmo raciocínio empregado com μ. Como $\hat{p}$ toma o lugar de $\bar{x}$, um intervalo de confiança aproximado de 95 por cento para p será dado pelas inequações

$$\hat{p} - 1{,}96\sqrt{\frac{pq}{n}} < p < \hat{p} + 1{,}96\sqrt{\frac{pq}{n}} \tag{6}$$

Substituindo-se p por $\hat{p}$ nos dois termos sob o radical e depois os valores de $n = 300$ e $\hat{p} = 0{,}12$, obtém-se o intervalo aproximado

$$0{,}12 - 0{,}037 < p < 0{,}12 + 0{,}037$$

Estes limites reduzem-se a 0,083 e 0,157; conseqüentemente um intervalo de confiança aproximado de 95 por cento para p será dado pelas desigualdades

$$0{,}083 < p < 0{,}157$$

A repetição pode ser enfadonha; mesmo assim vale a pena dizer mais uma vez que todas as três soluções são baseadas em métodos de grandes amostras. Felizmente estes métodos são muito bons, mesmo para pequenas amostras, desde que $np > 5$ para $p < 1/2$ e $nq > 5$ para $p > 1/2$. Há métodos melhores para valores pequenos de n, mas não serão tratados aqui

Exemplo. Uma ilustração final dos métodos para estimar p é dada visto ser de interesse para aqueles que gostam de política. Um pesquisador de opinião pública muito conhecido afirma que sua estimativa da proporção dos votantes favorecendo certo candidato presidencial não está em erro por mais de 0,03 unidades. Numa eleição presidencial muito equilibrada, que tamanho deveria ele tomar para a amostra a fim de estar certo, com uma probabilidade de 0,997, de que sua afirmativa é correta?

Da Tabela IV no apêndice, 99,7 por cento da área central de uma distribuição normal está compreendida dentro de três desvios-padrão da média; por conseguinte, $z_0 = 3$ aqui. Como a eleição é muito equilibrada, pode-se admitir que $p = 1/2$; assim, a fórmula (5) fornece o resultado

$$n = \frac{9 \cdot \frac{1}{2} \cdot \frac{1}{2}}{(0{,}03)^2} = 2\,500$$

Uma amostra aleatória deste tamanho tirada de todo o país deveria conseqüentemente ser suficiente para dar-lhe a precisão desejada.

O uso de $p = \frac{1}{2}$ no problema anterior pode parecer arbitrário, particularmente se a eleição não foi muito disputada. Contudo, é fácil demonstrar que pq assume seu valor máximo quando $p = \frac{1}{2}$, e dessa forma que o valor máximo de n na fórmula (5) ocorre quando $p = \frac{1}{2}$. Isto é feito verificando-se primeiro que

$$pq = p(1-p) = p - p^2 = \frac{1}{4} - \left(\frac{1}{2} - p\right)^2$$

Depois, pq será o maior possível, isto é, 1/4, quando o termo $(\frac{1}{2} - p)^2$ que está sendo subtraído de 1/4 tem o valor 0. Mas isto ocorrerá quando $p = \frac{1}{2}$. Isto implica que, quando se está

determinando o tamanho da amostra necessário para uma precisão especificada da estimativa, o valor de n para $p = \frac{1}{2}$ seja maior que qualquer outro valor de p. Como resultado, o uso de $p = \frac{1}{2}$ em tais problemas garante-nos que o valor resultante de n é certamente bastante grande e possivelmente maior que o necessário.

Uma característica interessante de problemas desse tipo é que, contrariamente à crença da maioria das pessoas, a precisão de uma estimativa de uma proporção p não depende do tamanho da população, mas só do tamanho da amostra. Dessa forma, uma amostra de 2 500 eleitores tirada de 50 000 000 de votantes é suficiente teoricamente para determinar a preferência desses eleitores, com alta precisão.

Infelizmente, eleitores não se comportam sempre como tentativas num jogo de azar, de modo que o modelo da distribuição binomial não é estritamente aplicável a problemas desse tipo. Por exemplo, um votante quando entrevistado pode favorecer um candidato e, contudo, uma semana mais tarde pode votar em outro candidato ou pode mesmo esquecer de votar. Pode também interpretar mal as perguntas do questionário, e assim responder de modo incorreto. A experiência tem mostrado que, devido a fatores humanos incontroláveis, a precisão de uma estimativa de p para votantes não aumenta muito após se tirar uma amostra de 10 000. É necessário bom senso na aplicação de modelos matemáticos à vida real, principalmente quando se trata de seres humanos e de suas inconsistências.

5 — A DISTRIBUIÇÃO t DE STUDENT

Os problemas da estimação de μ que foram resolvidos nas seções 2 e 3 supunham que o valor de σ fosse conhecido ou que a amostra fosse bastante grande para justificar a substituição de σ pela estimativa s de sua amostra. Como ilustração de um problema para o qual σ é desconhecido e para o qual métodos de grandes amostras seriam inadequados, suponhamos que um inspetor deseje inspecionar o peso do pão produzido numa padaria e toma uma amostra aleatória de 15 pães de uma fornada. Vamos supor que o peso médio e o desvio-padrão desses 15 pães sejam de 15,8 e 0,3 onças respectivamente, e que o inspetor queira um grau de confiança de 95 por cento para o peso médio da produção total. Esta amostra é, sem dúvida, pequena demais para fornecer uma boa estimativa de σ; portanto, para

evitar o erro envolvido na substituição de σ por s quando s está baseado numa tal amostra pequena, introduz-se uma nova variável chamada *variável t de Student*. Ela é definida pela fórmula

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} \tag{7}$$

Esta variável assemelha-se à variável normal padrão introduzida na seção 3, a saber:

$$z = \frac{\bar{x} - \mu}{\sigma}\sqrt{n}$$

Todavia difere de z por envolver o desvio-padrão da amostra s em lugar do desvio-padrão da população σ. Uma vez que t não requer um conhecimento de σ, como no caso de z, seu valor pode ser calculado dos dados da amostra, enquanto o valor de z não pode ser computado a menos que σ seja conhecido. Esta é a razão pela qual t pode ser usado para resolver problemas sem necessidade de se introduzir aproximações para os parâmetros de população.

Se fosse realizado um grande número de experimentos de amostragem nos quais fosse selecionada uma amostra de tamanho n de uma população normal e o valor de t fosse computado, um grande número de valores de t seria disponível para classificação numa tabela de freqüências a fim de obter-se uma boa estimativa da distribuição de t limite ou teórica. Os métodos matemáticos, contudo, conduzem à distribuição exata. Verifica-se que a distribuição de t depende apenas do valor de n, desde que a variável básica x possua uma distribuição normal. Além do mais, a distribuição de t é muito aproximada da distribuição de uma variável-padrão z, exceto para valores muito inferiores a n. A Figura 4 mostra o gráfico da distribuição de t para $n = 5$ e o gráfico para uma variável normal padrão z.

A Tabela V, no apêndice, dá os valores de t correspondentes aos vários "graus de liberdade" indicados por ν (*ni*), e várias probabilidades. Para o problema aqui considerado, o número de graus de liberdade é dado pela fórmula $\nu = n - 1$. Isto corresponde a usar o divisor $n - 1$ e não n na definição do desvio-padrão da amostra, no Capítulo 2. A distribuição t é usada também para outros tipos de problemas nos quais o parâmetro ν (*ni*) não é igual a n-1; de outro modo esta expressão não precisaria ser introduzida aqui. Qualquer cabeçalho de coluna, tal como 0,025, indica a probabilidade de t exceder numericamente o valor de t escrito nessa coluna. Por causa da simetria segue-se que 0,025 é também a probabilidade de que t esteja do lado esquerdo do valor correspondente $-t$.

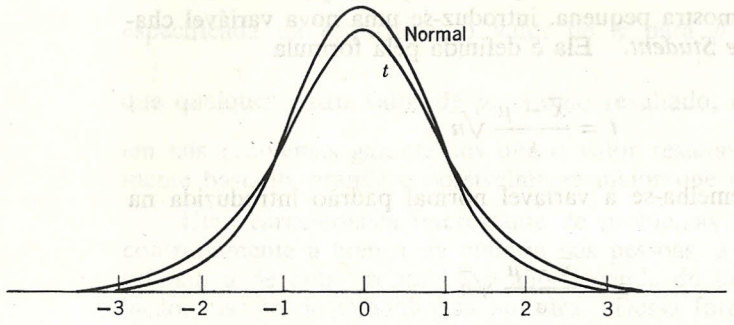


FIGURA 4. Distribuição normal padrão e distribuição t de Student quando n = 5.

Assim, para o problema acima, como $n = 15$ lê-se a entrada na linha correspondente a 14 graus de liberdade e na coluna encabeçada por 0,025 e se acha $t = 2{,}145$. Há, portanto, uma probabilidade de 0,95 de que t satisfaça as inequações

$$-2{,}145 < t < 2{,}145 \tag{8}$$

Se aplicarmos isso à fórmula (7), segue-se que a probabilidade é de 0,95 de que

$$-2{,}145 < \frac{\bar{x} - \mu}{s}\sqrt{n} < 2{,}145$$

Essas inequações podem ser resolvidas para μ da mesma maneira que pela fórmula (3). O resultado é

$$\bar{x} - 2{,}145\frac{s}{\sqrt{n}} < \mu < \bar{x} + 2{,}145\frac{s}{\sqrt{n}} \tag{9}$$

O desejado intervalo de confiança de 95 por cento para μ é obtido substituindo-se os valores da amostra $n = 15$, $\bar{x} = 15{,}8$ e $s = 0{,}3$ nessas inequações. O resultado destas substituições é

$$15{,}8 - 2{,}145\frac{0{,}3}{\sqrt{15}} < \mu < 15{,}8 + 2{,}145\frac{0{,}3}{\sqrt{15}}$$

Isto se simplifica em

$$15{,}63 < \mu < 15{,}97$$

Da análise precedente deveria aparecer que o peso médio dos pães desta padaria provavelmente é pouco menos que uma libra.

É importante perceber a distinção entre este novo método de encontrar o intervalo de confiança para μ e o método anterior da amostra maior. Este método não requer que o σ esteja próximo de s, como é verdade para o método da grande amostra, e conseqüentemente ele dá uma solução exata para o problema em lugar de uma solução aproximada.

A fórmula (9) não pode ser usada para encontrar o intervalo de confiança de 95 por cento para μ, a não ser que o tamanho da amostra seja 15, porque na Tabela V descobrimos que a fórmula (8) só vale para $v = 14$. Se t_0 for usado para significar o valor de t encontrado na coluna 0,025 da Tavela V em seguida a v (ni) graus de liberdade, então a fórmula geral para um intervalo de confiança de 95 por cento para μ é dado pelas inequações

$$\bar{x} - t_0\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_0\frac{s}{\sqrt{n}} \tag{10}$$

Esta fórmula é também válida para porcentagens outras que 95 se o valor correspondente de t_0 for empregado. Assim um intervalo de confiança de 90 por cento é obtido substituindo-se t_0 pelo valor de t da coluna 0,05 da Tabela V, o qual vem adiante do valor desejado dos graus de liberdade.

Como o método da pequena amostra baseado na distribuição t é um método exato, parece que sempre se deveria usá-lo quando σ fosse desconhecido. Contudo, a teoria que estuda a distribuição t requer que se suponha que a variável básica x possua uma distribuição normal; portanto a não ser que se esteja seguro de que x tenha, pelo menos, uma distribuição aproximadamente normal, a distribuição t pode não ser muito exata.

6 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Um aditivo para gasolina está sendo testado para ver se aumenta a quilometragem. Vinte e cinco carros recebem 5 galões de gasolina e são postos a andar até que a gasolina termine. No fim do experimento, calcula-se a quilometragem média para cada carro. Os cálculos forneceram uma média de $\bar{x} = 18{,}5$ milhas por galão e um desvio-padrão de $s = 2{,}2$ milhas por galão para os 25 carros. Experiências com carros da mesma espécie dos que foram usados antes não empregando aditivos indicou que, aproximadamente,

$\mu = 18{,}0$ e $\sigma = 2{,}0$ milhas por galão. Supondo que o aditivo não tenha nenhum efeito sobre a quilometragem, resolva os seguintes problemas: (a) 2 — Determine a precisão da probabilidade de $\bar{x}$ como uma estimativa de μ. Qual é a precisão real? A amostra é compatível com o que é esperado pela teoria? (b)2 — Mostre o tamanho do experimento a ser feito se quisermos ter a certeza com um grau de probabilidade de 0,95 de que a experiência não comete erros de mais de 1/2 milha por galão. (c) 2 — Encontre um intervalo de confiança de 95 por cento para μ. Esse intervalo contém realmente μ? (d) 5 — Deixando de supor que o aditivo não tenha nenhum efeito sobre a média ou a variância, use a variável t de Student para encontrar um intervalo de confiança de 95 por cento para μ. Seguem-se as soluções.

(a) Aqui $\mu = 18{,}0$, $\sigma = 2{,}0$, $n = 25$, e $\bar{x} = 18{,}5$; dessa forma

$$1{,}96\ \sigma_{\bar{x}} = 1{,}96\ \frac{2{,}0}{\sqrt{25}} = 0{,}784$$

Há uma probabilidade de 0,95 de que o erro de estimativa não exceda a 0,784. O erro real é $|\bar{x} - \mu| = 18{,}5 - 18{,}0 = 0{,}5$; portanto, o valor amostral é compatível com a teoria.

(b) Como 0,95 dá $z_0 = 1{,}96$ e o erro máximo tolerável deve ser $e = \frac{1}{2}$, n é dado por

$$n = \left[\frac{(1{,}96)(2{,}0)}{\frac{1}{2}}\right]^2 = 61{,}5$$

Uma amostra adicional de 37 bastaria. Desta forma, deveriam ser empregados, aproximadamente, 37 carros adicionais do mesmo tipo.

(c) Um intervalo de confiança de 95 por cento é dado por

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Dessa forma, para este problema torna-se

$$18{,}5 - 1{,}96\frac{2{,}0}{\sqrt{25}} < \mu < 18{,}5 + 1{,}96\frac{2{,}0}{\sqrt{25}}$$

ou

$$17{,}72 < \mu < 19{,}28$$

Como $\mu = 18{,}0$, está contido dentro do intervalo.

(d) Como $\nu = 24$ e se deseja um intervalo de 95 por cento, $t_0 = 2{,}0639$; assim a fórmula (10) nos dá

$$18{,}5 - 2{,}06\frac{2{,}2}{\sqrt{25}} < \mu < 18{,}5 + 2{,}06\frac{2{,}2}{\sqrt{25}}$$

ou

$$17{,}59 < \mu < 19{,}41$$

Este intervalo também contém $\mu = 18{,}0$; portanto é compatível com o pressuposto do não-efeito do aditivo.

Exemplo 2. Uma companhia de seguros descobriu por experiência que 30 por cento de seus segurados tinham tido pelo menos um acidente automobilístico nos três últimos anos passados. Está pretendendo assegurar todos os empregados do serviço civil de certa cidade e deseja determinar sua taxa de segurança. Toma uma amostra de 100 desses empregados e descobre que 25 deles tiveram pelo menos um acidente nos últimos três anos. Supondo que estes empregados sejam típicos dos segurados da companhia, faça os problemas seguintes: (a) 4 — Qual a precisão da probabilidade desta estimativa? Qual é a precisão real? A proporção da amostra é compatível com aquilo que deveríamos esperar pela teoria? (b) 4 — Ache o tamanho necessário de uma amostra adicional, se quisermos estimar p dentro de 0,03 unidades com uma probabilidade de acerto de 0,95. Suponha que p não seja conhecido aqui. (c) 4 — Solucione a parte (b) se o valor conservador $p = \frac{1}{2}$ for usado no lugar da estimativa amostral de p. (d) 4 — Encontre um intervalo de confiança de 90 por cento para p, supondo que o valor p seja desconhecido. Este intervalo realmente contém p? Eis as soluções.

(a) Aqui $\hat{p} = 0{,}25$. Como $p = 0{,}30$ e $n = 100$

$$1{,}96\ \sigma_{\hat{p}} = 1{,}96\ \sqrt{\frac{(0{,}30)\ (0{,}70)}{100}} = 0{,}09$$

Há uma probabilidade de 0,95 de que o erro não exceda a 0,09. O erro real é $|\hat{p} - \mathrm{p}| = |0{,}25 - 0{,}30| = 0{,}05$; portanto o valor da amostra é compatível com a teoria.

(b) Como $z_0 = 1{,}96$ e $e = 0{,}03$, n é dado por

$$n = \frac{(1{,}96)^2\ pq}{(0{,}03)^2} = \frac{(1{,}96)^2\ (0{,}25)\ (0{,}75)}{(0{,}03)^2} = 800$$

Dessa forma, seria preciso uma amostra adicional de 700.

(c) Usando $p = \frac{1}{2}$,

$$n = \frac{(1{,}96)^2\ (\tfrac{1}{2})\ (\tfrac{1}{2})}{(0{,}03)^2} = 1067$$

Assim, necessitar-se-ia de 967 amostras adicionais. Usando este valor muito conservador de p o custo da amostragem aumentaria muito aqui.

(d) usando $\hat{p} = 0{,}25$ em lugar de p sob o radical da fórmula (6) teremos

$$0{,}25 - 1{,}64\sqrt{\frac{(0{,}25)\ (0{,}75)}{100}} < p < 0{,}25 + 1{,}64\sqrt{\frac{(0{,}25)\ (0{,}75)}{100}}$$

ou

$$0{,}18 < p < 0{,}32$$

O valor $p = 0{,}30$ está dentro deste intervalo.

Exemplo 3. Um artigo no **Canadian Medical Association Journal,** de setembro de 1972, discute os efeitos de fortes doses de vitamina C sobre os resfriados. Conclui, entre vários resultados, que a vitamina C é um tanto benéfica na redução do tempo requerido para a recuperação do resfriado. Neste experimento, 407 indivíduos que tomaram fortes doses de vitamina C e que tinham contraído resfriado levaram em média 5,25 dias para a recuperação com um desvio-padrão de 6,0 dias. Este grande desvio-padrão é devido, sem dúvida, ao fato de alguns indivíduos levarem muito tempo para se recuperar de um resfriado e, portanto, a distribuição de tempos tem uma longa cauda à direita. 3 — Com estes dados, calcule um intervalo de confiança de 90 por cento para a média da distribuição do tempo de recuperação.

Como $n = 407$ é uma amostra grande, basta substituir σ por s na fórmula

$$\bar{x} - 1{,}64\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}64\frac{\sigma}{\sqrt{n}}$$

Substituindo os valores da amosta $\bar{x} = 5{,}25$, $s = 6{,}0$ e $n = 407$ temos

$$5{,}25 - 1{,}64\frac{6{,}0}{\sqrt{407}} < \mu < 5{,}25 + 1{,}64\frac{6{,}0}{\sqrt{407}}$$

Isto se reduz a

$$4{,}76 < \mu < 5{,}74$$

Neste experimento, a média para o grupo que não tomou doses de vitamina C foi de 6,02 dias.

Exemplo 4. Um artigo no jornal médico *Lancet* de 2 de março de 1974 discute os efeitos dos anticoncepcionais orais sobre a razão dos sexos, no nascimento. Num estudo levado a cabo na Hungria, registraram-se os sexos das crianças nascidas de mães que tinham tomado a "pílula" durante um espaço de tempo de pelo menos 2 anos antes do nascimento da criança. Havia 170 mães, nessa circunstância, no experimento. O número de crianças do sexo masculino nascidos neste grupo foi de 58 e, portanto, o número de crianças do sexo feminino foi de 112. 4 — Use os dados para calcular um intervalo de confiança de 90 por cento para a percentagem de crianças do sexo masculino esperada desse tipo de mães.

Podemos aplicar, aqui, a fórmula (6) substituindo 1,96 por 1,64. A substituição dos valores da amostra $n = 170$ *e* $\hat{p} = \frac{58}{170} = 0{,}34$ nessa fórmula dá

$$0{,}34 - 1{,}64\sqrt{\frac{(0{,}34)\ (0{,}66)}{170}} < p < 0{,}34 + 1{,}64\sqrt{\frac{(0{,}34)\ (0{,}66)}{170}}$$

Isso se reduz a

$$0{,}28 < p < 0{,}40$$

Deste estudo se infere que há, aproximadamente, duas vezes mais crianças do sexo feminino que do sexo masculino nascidas de mães que tomaram anticoncepcionais durante pelo menos 2 anos antes do nascimento da criança.

EXERCÍCIOS

Seção 2

1 — A experiência com trabalhadores de certa indústria indica que o tempo requerido para que um trabalhador aleatoriamente selecionado realize um serviço é distribuído de maneira aproximadamente normal como um desvio-padrão de 12 minutos. (a) Se cada tra-

balhador de uma amostra aleatória de 16 realizasse o trabalho, que grau de precisão deveria ter a média da amostra como estimativa da média para todos os trabalhadores? (b) Que aumento teria havido na precisão dessa estimativa se tivessem sido selecionados 64 trabalhadores? (Use a probabilidade de 0,95.)

2 — Da experiência passada sabe-se que o desvio-padrão da altura de crianças do quinto ano num sistema escolar é de 2 polegadas. (a) Se fosse tomada uma amostra aleatória de 25 de tais crianças, de que precisão seria a média da amostra como estimativa da média de todas as crianças? (b) O que sucederia à precisão desta estimativa se a amostra fosse 4 vezes maior? Use uma probabilidade de 0,95.

3 — (a) No problema 1(a), de que tamanho alguém precisaria tomar uma amostra a fim de estimar a média da população com um erro máximo de 3 minutos com uma probabilidade de 0,95 de estar correto? (b) Que tamanho de amostra seria necessário se o máximo erro de estimativa fosse de 1 minuto?

4 — Faça o problema 3 para o caso no qual se fique satisfeito com a probabilidade de 0,90 de estar correto.

5 — (a) No problema 2, que tamanho se precisaria tomar para a amostra a fim de estimar a média da população com um erro máximo de 1/2 polegada, com uma probabilidade de 0,95 de se estar correto? (b) Que tamanho de amostra seria necessário se o erro máximo da estimativa tivesse de ser de 1/5 de polegada?

6 — Faça o problema 5(a) para o caso no qual se fique satisfeito com a probabilidade de 0,90 de estar certo.

7 — Se os resultados do experimento no problema 1(a) dessem $\bar{x} = 140$ minutos, encontrar (a) limites de confiança de 95 por cento para μ, (b) limites de confiança de 90 por cento para μ.

8 — Se os resultados do experimento do problema 2(a) tenham dado $\bar{x} = 54$ polegadas, encontrar (a) limites de confiança de 95 por cento para μ, (b) limites de confiança de 90 por cento para μ.

Seção 3

9 — Um conjunto de 60 animais experimentais é alimentado com certo tipo de ração por um período de 2 semanas. Seus aumentos de peso deram os valores de $\bar{x} = 42$ onças e $s = 4$ onças. (a) Qual a precisão de 42 como estimativa da média da população? (b) De que tamanho você tomaria uma amostra se desejasse que o $\bar{x}$ diferisse de μ por uma quantidade menor que $\frac{1}{2}$ onça, com uma probabilidade de 0,95 de estar correto? (c) Encontre limites de confiança de 95 por cento para μ. Use aqui os métodos das grandes amostras.

10 — Se se tomar uma amostra de 100 e resultar $\bar{x} = 40$, $s = 6$, com que probabilidade podemos estar certos de que $\bar{x}$ não se afasta da média em mais de 1 unidade?

11 — Para se estimar a quilometragem de certo modelo de carro, foram selecionados 30 carros do mesmo modelo e testados num percurso bastante longo. Se a média e o desvio-padrão de suas quilometragens forem de 19,6 e 0,7, respectivamente, encontre um intervalo de confiança de 90 por cento para a quilometragem média deste tipo de carro. Suponha que a quilometragem seja uma variável normal e que grandes métodos de amostra são satisfatórios aqui.

12 — Um criador de frangos tem um lote de 1 000 animais. Deseja experimentar uma nova dieta para aumento de peso de seu lote. Se ele julga que é de 2 onças o desvio-padrão do aumento do ganho de peso, num período de 1 mês, qual o tamanho da amostra que ele devia tomar para fins experimentais, de modo que sua estimativa do ganho total em peso para os 1 000 frangos submetidos à dieta não fosse errada em mais de 40 libras, com uma probabilidade de 0,95?

13 — No problema 12 o valor do desvio-padrão usado baseava-se no palpite do criador; poderia, portanto, ser pouco seguro. Sugira como você determinaria com maior precisão o tamanho da amostra necessária, tirando amostras em pequenos grupos e reestimando σ como se se tomassem grupos de amostra adicional.

Seção 4

14 — Certa organização de um campus deseja estimar a percentagem de estudantes que favorecem a constituição de um novo corpo estudantil. Propõe selecionar uma amostra aleatória de 300 estudantes. Se o resultado do inquérito fornecer $\hat{p} = 0{,}60$, qual a precisão provável desta estimativa com relação à verdadeira proporção? Use uma probabilidade de 0,95.

15 — Se se desejar para o problema 14 uma estimativa precisa em 0,04 unidades para **p**, qual seria o tamanho da amostra que a organização deveria tomar? Considere como suficiente uma probabilidade de exatidão de 0,95 e use $\hat{p} = 0{,}60$.

16 — Um fabricante de peças julga que aproximadamente 3 por cento de seu produto contém defeitos. Se deseja estimar a verdadeira percentagem dentro de $\frac{1}{2}$ por cento e estar seguro da correção com uma probabilidade de 0,97, que tamanho da amostra deve ser tomado?

17 — Se a organização do campus do problema 15 não tivesse nenhuma experiência para fazer a estimativa $\hat{p} = 0{,}60$, que tamanho de amostra deveria tomar?

18 — Contrata-se um profissional para estimar que proporções de votantes favorecerá determinada questão. Se a estimativa tiver de cair dentro das 0,03 unidades da verdadeira proporção com uma probabilidade de 0,90, que tamanho de amostra ele deve tomar?

19 — Uma amostra aleatória de 400 cidadãos de uma comunidade mostrou que 280 queriam ter sua água fluoretada. Use estes dados para encontrar limites de confiança de 95 por cento para a proporção da população que favorece a fluoretação.

20 — Num inquérito efetuado pelo Dr. McGuire, da Escola de Medicina da Universidade de Irvine, Califórnia, e relatado no **Los Angeles Times** de 11 de julho de 1974, descobriu que pouco menos de 40 por cento dos 2 000 motoristas mortos em acidentes automobilísticos durante os últimos 5 anos, na comarca de Orange, estavam embriagados. A definição de embriaguez dada pelo California Vehicle Code baseia-se na quantidade de álcool no sangue, e a determinação da embriaguez ou não do motorista funda-se num teste **post mortem** de seu sangue. Na pressuposição de que este período de 5 anos seja típico da experiência motorística da comarca de Orange (e provavelmente valerá também para o futuro), calcule um intervalo de confiança de 95 por cento para a percentagem de motoristas de Orange que morrerão em estado de embriaguez.

21 — Certa organização de um campus universitário desejava estimar a percentagem do corpo discente que favorecesse o seu candidato numa próxima eleição. Para fazer isso, interrogou os 200 primeiros estudantes, que se dirigiam às aulas das 8 horas, sobre suas opiniões. Descobriram que 30 por cento deles favoreciam seu candidato. (a) Critique a validade desta estimativa. (b) Se a estimativa fosse válida, quais seriam os limites de confiança de 96 por cento para **p**?

22 — Obter limites de confiança de 90 por cento para o número de reclamações de acidentes que serão pagos por uma seguradora durante o próximo ano se a experiência deste ano mostrou que houve 7 por cento de tais reclamações e a seguradora tem 7 000 apólices.

23 — Resolva cada uma das desigualdades da dupla inequação $\hat{p}-z_0\sqrt{pq/n} < p < \hat{p}+z_0\sqrt{pq/n}$ para a variável **p**. Isto envolverá a solução de uma equação quadrática. Use seus resultados para obter um intervalo de confiança para **p** que não contenha **p** nos seus limites.

Seção 5

24 — Dado que $\bar{x} = 20$, $s = 4$, $n = 12$, com **x** normalmente distribuído, use a distribuição **t** de Student para encontrar (a) limites de confiança de 95 por cento para μ, (b) limites de confiança de 99 por cento para μ.

25 — Uma amostra de 20 cigarros de certa marca foi testada quanto ao conteúdo de nicotina e deu $\bar{x} = 22$ e $s = 4$ miligramas. Use a distribuição **t** de Student para encontrar limites de confiança de 95 por cento para μ.

26 — Resolva o problema 25 utilizando os métodos de grandes amostras e compare os resultados dos dois métodos.

27 — Um conjunto de 12 animais experimentais recebeu uma dieta especial por 3 semanas e conseguiu os seguintes aumentos de peso: 30, 22, 32, 26, 24, 40, 34, 36, 32, 33, 28 e 30. Ache limites de confiança de 90 por cento para μ.

28 — Administrou-se um sonífero a 20 pacientes e houve um aumento de 32 minutos na média do sono dessas pessoas. Se o desvio-padrão dos 20 pacientes for de 18 minutos, encontre um intervalo de confiança de 90 por cento para o aumento da média dos pacientes que estão em tratamento.

29 — No problema 28 ache os intervalos de confiança que teriam resultado se a amostra tivesse sido do tamanho de (a) 10, (b) 40. Com base nesses dois resultados, o que se pode dizer sobre a amplitude do intervalo de confiança como função do tamanho da amostra?

Geral

30 — Os dados seguintes foram retirados de um estudo realizado por vários médicos e aparece no **New England Journal of Medicine** de abril de 1972. Neste estudo fez-se uma análise do grau do enfisema presente em 55 homens que nunca tinham fumado, 31 homens que fumavam cachimbo ou charuto e 401 que fumavam cigarros. Todos tinham menos de 60 anos. Os fumantes mais antigos foram considerados num estudo separado. A escala desta tabela difere um pouquinho da escala da tabela do artigo original.

Grau do Enfisema	Não Fumantes	Fumantes de Cachimbo ou Charuto	Fumantes de Cigarr		
			<1	1-2	>2
0-1	53	18	15	2	0
1-2	2	11	13	24	5
2-3		1	19	130	56
3-4		1	17	50	38
4-5			4	8	7
5-7				4	5
7-9				3	1
Totais	55	31	68	221	112

Seja **x** a variável que mede o grau de enfisema, e usando os valores $x_1 = 0{,}5$, $x_2 = 1{,}5$, $x_3 = 2{,}5$, $x_4 = 3{,}5$, $x_5 = 4{,}5$, $x_6 = 6{,}0$, $x_7 = 8{,}0$, (a) calcule a média e o desvio-padrão de **x** para o grupo de fumantes de cigarro, combinando os dados numa única tabela; (b) calcule um intervalo de confiança de 95 por cento para a média dos fumantes baseado nos cálculos de (a); (c) calcule a média do grupo dos não fumantes e dos fumantes de cachimbo e charuto e observe se eles caem no intervalo obtido em (b).

31 — Um estudo feito por três professores de psiquiatria do Department of Psychiatry, Rutgers Medical School, relatado na publicação de fevereiro de 1974 dos **Archives of General Psychiatry**, apresentou os seguintes dados sobre as atitudes dos United Automobile Workers que trabalhavam numa fábrica da General Motors.

Qual o grau de satisfação com seu emprego?	Pacientes Ativos	Primeiros Pacientes	Satisfeitos classificados	Aborrecidos classificados
Número de satisfeitos	13	19	90	463
Tamanho da amostra	17	26	95	481
Perçentagem	76	73	95	96

Os dois grupos de "pacientes" consistem nos trabalhadores que foram vistos e diagnosticados na UAW Labor Union Clinic no Johns Hopkins Hospital. Os restantes foram classificados como "doentes" ou "bons" na base do MacMillan Index, um estratagema criado para medir a saúde mental por meio de entrevistas.

(a) Ache um intervalo de confiança de 95 por cento para a percentagem dos trabalhadores bons que estão satisfeitos com seu trabalho.

(b) Ache um intervalo de confiança de 90 por cento para a percentagem de trabalhadores doentes satisfeitos com seu emprego.

(c) Com base nestes dados é verdade que a satisfação do trabalhador com relação ao emprego é independente das taxas altas obtidas na entrevista sobre a saúde mental, se ele não foi paciente na Clínica?

32 — Faça cada membro da classe encontrar um intervalo de confiança de 75 por cento para μ para uma amostra de tamanho 25, numa tabela de números aleatórios de um dígito (Tabela XI do apêndice). Usar o fato de σ para essa distribuição ser dado como $\sigma = 2{,}87$ e $\bar{x}$ ser tratado como variável normal. Testar para ver que percentagem dos intervalos de confiança dos estudantes contém a média verdadeira $\mu = 4{,}5$. Cerca de 75 por cento deles deveriam satisfazer a esta condição.

33 — Faça o problema 32, mas desta vez use uma amostra de tamanho 10 e suponha que o valor de σ não é conhecido. Isto é, use a distribuição **t** de Student para encontrar o intervalo de confiança desejado. Testar para ver que percentagem dos intervalos dos estudantes contém μ.

34 — Os dados do problema 28, Capítulo 2, podem ser usados para problemas de estimação, como a estimação da média de um conjunto e a determinação de sua precisão.

35 — Os dados da Tabela XII no apêndice podem ser usados para construir vários problemas de estimativa. Por exemplo, qual a segurança dada pela média da variável $\mathbf{x}_5$ como estimativa da média da população estudantil para esta variável?

8

TESTES DE HIPÓTESE

1 — DOIS TIPOS DE ERRO

Conforme indicado no Capítulo 1, um segundo problema fundamental de estatística é o teste de hipótese sobre as populações. Da discussão sobre estimativa no Capítulo 7, segue-se que o teste de hihipóteses sobre populações binomiais pode usualmente ser reduzido ao teste de alguma hipótese sobre o parâmetro p. De modo análogo, o teste de hipóteses sobre populações normais pode usualmente ser reduzido ao teste de hipóteses sobre os parâmetros μ e σ.

Já foram discutidos exemplos que são essencialmente problemas de teste de hipóteses relacionados com distribuições binomiais ou normais. Por exemplo, o problema de determinar por meio de uma amostra de 400 eleitores, se era verdadeira a afirmativa de um político sobre 60 por cento de apoio, é um problema de testar a hipótese de que $p = 0{,}6$ para a distribuição binominal em que $n = 400$. O problema de comparar os pesos dos estudantes internos e externos apresentado no Capítulo 2 pode ser tratado como um problema de testar a hipótese da igualdade das médias e dos desvios-padrão para as duas distribuições normais.

Com o fim de explicar os métodos usados para testar uma hipótese sobre o parâmetro de uma distribuição, consideremos um problema particular.

Durante os últimos 50 anos ou mais, os arqueólogos de certo país têm tentado classificar os crânios encontrados em escavações em dois grupos raciais, em parte pela cerâmica e outros utensílios em contrados com os crânios e em parte pelas diferenças nas dimensões

destes. Em especial, eles descobriram que o comprimento médio de todos os crânios encontrados até aqui afasta-se da média da raça A em 190 milímetros ao passo que para a raça B o afastamento é de 196 milímetros. O desvio-padrão de tais medidas de comprimento foi constatado como aproximadamente o mesmo para os dois grupos e igual a cerca de 8 milímetros. Uma nova escavação forneceu 12 crânios, os quais podem ser considerados, com certa razão, como pertencentes à raça A. O comprimento da média destes crânios é $\bar{x} = 194$ milímetros. O problema é testar a hipótese de que os crânios pertençam à raça A de preferência à B.

Como o teste deve ser baseado no valor de $\bar{x}$ ele é formulado como um teste de hipótese, indicado por H_0, de que a média da população para os 12 crânios seja de 190 em contraste com a hipótese alternativa, indicada por H_1, de que a média da população seja de 196. Isto pode ser condensado como segue:

$$(1) \qquad \begin{aligned} H_0 &: \mu = 190 \\ H_1 &: \mu = 196 \end{aligned}$$

Há aqui duas possibilidades de se tomar decisões erradas. Se os crânios realmente pertencem à raça A e com base no valor de $\bar{x}$ se decide aceitar H_1, está sendo tomada uma decisão incorreta. Se, todavia, os crânios realmente pertencem à raça B e se decide aceitar H_0, também está sendo tomada uma decisão incorreta. O primeiro tipo de decisão errada é geralmente chamado erro do tipo I e o segundo tipo de decisão errada é chamado erro do tipo II. Estas duas possibilidades de decisões incorretas, juntamente com as duas possibilidades de decisões corretas, estão enumeradas no Quadro 1.

QUADRO 1

	Aceita-se H_0	Aceita-se H_1
H_0 verdadeira	decisão correta	erro do tipo I
H_1 verdadeira	erro do tipo II	decisão correta

A maioria das pessoas usaria o bom senso neste problema particular e decidiria em favor de H_0 se $\bar{x}$ estivesse mais próximo de 190 do que de 196 e em favor de H_1 se o inverso se verificasse. Assim, a maior parte das pessoas aceitaria H_1 neste problema. Contudo, os arqueólogos que têm outras razões para acreditar que os crânios pertençam à raça A, tais como peças de cerâmica encontradas com eles, não desejariam usar o ponto a meio caminho entre as duas médias como valor-limite para tomar decisões baseadas em $\bar{x}$. Eles sem dúvida insistiriam em que $\bar{x}$ fosse claramente próximo da média correspondente a H_1 antes de quererem abandonar a hipótese H_0 em favor de H_1. A fim de estudar a racionalidade de usar o ponto do meio ou outros pontos à sua direita, para tomar decisões, serão calculadas as probabilidades de fazer os dois tipos de erro.

Com o propósito de calcular estas probabilidades, admite-se que x, o comprimento do crânio, tenha distribuição aproximadamente normal, com desvio-padrão $\sigma = 8$ e com média $\mu = 190$ se o crânio for da raça A, com a média $\mu = 196$ se o crânio for da raça B. Então $\bar{x}$ pode ser suposta como normalmente distribuída com desvio-padrão

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{12}} = 2{,}31$$

e média 190 se os crânios forem da raça A e média 196 se forem da raça B. Os gráficos das duas curvas normais para $\bar{x}$ correspondendo a H_0 e H_1 são apresentados na Figura 1.

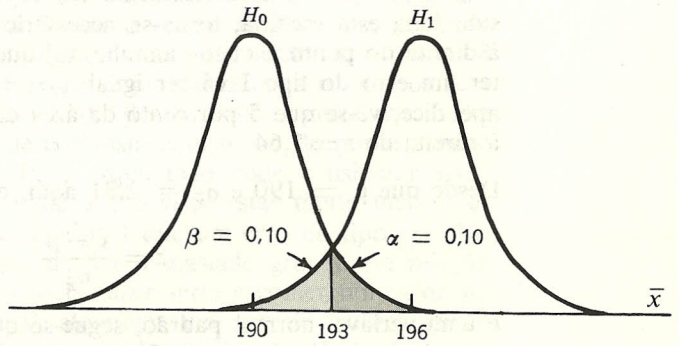


FIGURA 1. Distribuição de $\bar{x}$ sob H_0 e H_1.

Se o ponto a meio caminho, 193, for usado para limite das decisões, então a probabilidade de se cometer um erro do tipo I, isto é, a probabilidade de se aceitar H_1 quando H_0 é verdadeira é a probabilidade de que $\bar{x} > 193$ quando H_0 é verdadeira. Esta probabilidade é igual à área hachurada sob a curva de H_0 à direita de $\bar{x} = 193$. Seu valor, que é indicado por α, foi encontrado como

sendo 0,10, por método tipo II, isto é, a probabilidade de se aceitar H_0 quando H_1 for verdadeira é a probabilidade de que $\bar{x} < 193$ quando H_1 é verdadeira. Esta probabilidade é igual à área sombreada sob a curva de H_1 à esquerda de $\bar{x} = 193$. Seu valor, simbolizado por β, é, por simetria, o mesmo de α; dessa forma $\alpha = 0{,}10$ e $\beta = 0{,}10$.

Se o arqueólogo está bastante confiante, através de outras fontes de informação, de que a hipótese H_0 é verdadeira, ele desejará tomar a probabilidade de rejeitar H_0 sendo ela realmente verdadeira, consideravelmente menor que a probabilidade de rejeitar H_1 sendo ela realmente verdadeira. Assim, ele quererá que α seja consideravelmente menor que β. A Figura deixa claro que se fosse escolhido para limite das decisões um ponto à direita de 193, o valor de α se tornaria menor que 0,10 e o valor de β se tornaria maior que 0,10. Como não é possível diminuir α sem aumentar β, o arqueólogo deverá sentir algum constrangimento em diminuir α sem o que ele se defrontará com o valor de β demasiado grande. Suponha que ele decida que um valor de $\alpha = 0{,}05$ será pequeno o suficiente para lhe assegurar a proteção desejada contra o fato de rejeitar H_0 incorretamente. Significa que em apenas cerca de um experimento em vinte ele rejeitará incorretamente H_0 sendo ela verdadeira. Tendo sido feita esta escolha, torna-se necessário selecionar um valor de $\bar{x}$, à direita do ponto a meio caminho, tal que a probabilidade de cometer um erro do tipo I vá ser igual a $\alpha = 0{,}05$. Pela Tabela IV no apêndice, vê-se que 5 por cento da área da curva normal-padrão fica à direita de $z = 1{,}64$.

Desde que $\mu = 190$ e $\sigma_{\bar{x}} = 2{,}31$ aqui, e como

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

é uma variável normal padrão, segue-se que o valor de $\bar{x}$ que separa uma extremidade direita de 5 por cento da curva de $\bar{x}$ é obtido resolvendo-se a equação em $\bar{x}$.

$$1{,}64 = \frac{\bar{x} - 190}{2{,}31}$$

A solução desta equação, que é indicada por $\bar{x}_0$, é dada por $\bar{x}_0 = 193{,}8$. Outra maneira de chegarmos a este valor é mostrar que é necessário ir 1,64 desvios-padrão para a direita da média da distribuição normal para se obter um valor tal que 5 por certo da área sob a curva esteja à direita dela. Assim, $\bar{x}_0$ deve ser dado por

$$\bar{x}_0 = 190 + 1{,}64(2{,}31) = 193{,}8$$

Assim, segue-se que H_1 deveria ser aceita aqui, porque o valor da amostra $\bar{x} = 194$ está à direita de $\bar{x}_0 = 193{,}8$.

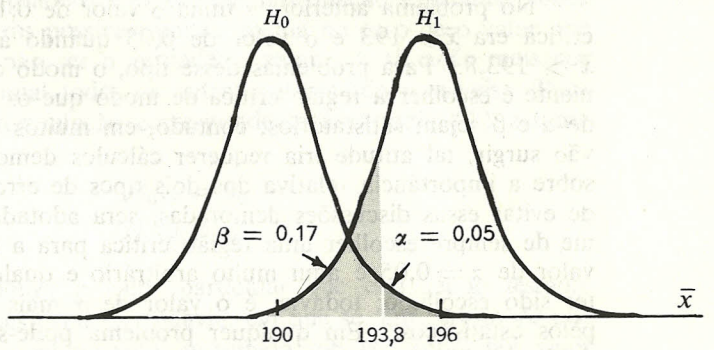


FIGURA 2. Distribuição de $\bar{X}$ sob H_0 e H_1, com região crítica selecionada.

Com esta escolha de $\bar{x}$ como o valor-limite para tomar decisões, o valor de β fica sendo a área sob a curva de H_1 à esquerda de $\bar{x} = 193{,}8$. Pelos métodos explanados no Capítulo 4, o valor β encontrado será de 0,17. A figura 2 apresenta estes resultados geometricamente. Embora o valor de β seja aqui consideravelmente maior que o valor de α, o que contrasta com o uso do ponto a meio caminho tornou $\beta = \alpha = 0{,}10$, o arqueólogo pode considerar satisfatórios os tamanhos relativos de α e β visto estar muito mais interessado em cometer um erro do tipo I que um erro do tipo II. No caso de ele sentir que o valor de β é demasiado grande em relação ao valor de α, tudo que precisaria fazer seria escolher um valor suficientemente pequeno para $\bar{x}_0$ até obter um par de valores que o satisfizesse. A parte do eixo dos x à direita de $\bar{x}_0$ é chamada *região crítica* do teste. Consiste naqueles valores de $\bar{x}$ correspondentes à rejeição de H_0. O método para testar a hipótese H_0 por meio de $\bar{x}$ pode ser expresso muito simplesmente em termos de sua região crítica estabelecendo-se que a hipótese H_0 será rejeitada se o valor $\bar{x}$ da amostra cair na região crítica do teste; caso contrário, H_0 será aceita.

Este método de testar a hipótese H_0 é o método que será usado neste livro para o teste de várias hipóteses. Consiste essencialmente em selecionar uma região crítica para a variável que está em jogo a

fim de testar a hipótese de que a probabilidade da variável que cai na região crítica seja um valor fixo α, concordando então em rejeitar esta hipótese se, somente se, o valor da amostra da variável cair na crítica. O valor experimental de $\bar{x}$ é usado apenas para tomar uma decisão após ter sido selecionada a região crítica, não se permitindo nunca que se influencie na seleção dessa região.

No problema anterior, α tinha o valor de 0,10 quando a região crítica era $\bar{x} > 193$ e o valor de 0,05 quando a região crítica era $\bar{x} > 193{,}8$. Para problemas desse tipo, o modo de proceder conveniente é escolher a região crítica de modo que os tamanhos relativos de α e β sejam satisfatórios; contudo, em muitos dos problemas que vão surgir, tal atitude iria requerer cálculos demorados e discussões sobre a importância relativa dos dois tipos de erro em jogo. A fim de evitar essas discussões demoradas, será adotada a atitude uniforme de sempre escolher uma região crítica para a qual $\alpha = 0{,}05$. O valor de $\alpha = 0{,}05$ é aqui muito arbitrário e qualquer outro poderia ter sido escolhido; todavia, é o valor de α mais comumente usado pelos estatísticos. Em qualquer problema pode-se calcular o valor de β e então ajustar o valor de α no caso de o valor de β não ser satisfatório quando $\alpha = 0{,}05$. Isto funciona para ambos os casos, naturalmente. Para um experimento muito grande, com α fixado em 0,05, pode acontecer que β seja consideravelmente menor que 0,05. Se o erro do tipo I for considerado mais sério que um erro do tipo II, será preciso ajustar o teste a fim de tornar α menor que β, o que então certamente tornaria α menor que 0,05.

Este método de testar hipóteses requer que se escolha uma região crítica para a qual $\alpha = 0{,}05$, mas por outro lado não determina como vai ser escolhida essa região crítica. Uma vez que é desejável tornar tão pequenas quanto possível as probabilidades dos dois tipos de erro e visto que α foi fixado, deve-se escolher uma região crítica que torne β tão pequeno quanto possível. Embora a escolha da região crítica na Figura 2 fosse baseada no bom senso, com a restrição de que $\alpha = 0{,}05$ pode ser demonstrado que nenhuma outra região crítica com $\alpha = 0{,}05$ terá um valor de β tão pequeno quanto o de $\beta = 0{,}17$.

Felizmente, na maior parte dos problemas simples, o bom senso de um indivíduo ou sua intuição o levará à escolha de uma região que seja a melhor possível, para minimizar o valor de β. Para problemas mais difíceis no teste de hipóteses, há uma teoria matemática que permite aos estatísticos encontrar regiões críticas melhores. As regiões críticas escolhidas nos problemas a serem resolvidos neste capítulo e em capítulos posteriores são as obtidas pelo uso desta teoria sempre que ela se aplique.

2 — TESTE DE MÉDIA

O problema discutido na seção 1 é uma ilustração do problema geral de testar a hipótese de que a média de uma população normal particular tenha certo valor. Aquele problema era muito incomum pelo fato de só haver um valor alternativo para a média. Na maioria dos problemas práticos não se tem informação específica sobre os valores alternativos possíveis para a média no caso de o valor que está sendo testado não ser o verdadeiro valor. A situação mais comum é aquela na qual todos os outros valores são possíveis. Para tais problemas, a formulação correspondente à fórmula (1) assume a forma

$$\begin{aligned} &H_0: \mu = \mu_0 \\ &H_1: \mu \neq \mu_0 \end{aligned} \tag{2}$$

Aqui μ_0 simboliza o valor particular que está sendo testado. Há muitos problemas práticos, todavia, em que se tem certeza de que se a média não é igual ao valor postulado sob H_0, então seu valor deve ser maior que esse valor postulado. Para tais problemas a fórmula (2) deveria ser substituída por

$$\begin{aligned} &H_0: \mu = \mu_0 \\ &H_1: \mu > \mu_0 \end{aligned} \tag{3}$$

Para problemas em que se esteja bem certo de que se a média não é igual a μ_0 devendo então seu valor ser menor que μ_0, dever-se-á, é claro, substituir $\mu > \mu_0$ por $\mu < \mu_0$ na fórmula (3).

Exemplo 1. Como ilustração, suponhamos que uma cidade venha comprando a marca **A** de lâmpadas por vários anos mas esteja pensando em trocá-la pela marca **B** por causa do melhor preço. Os vendedores da marca **B** afirmam que seu produto é tão bom quanto o da marca **A**. A experiência de vários anos tem demonstrado que as lâmpadas da marca **A** têm uma vida média de 1 180 horas, com um desvio-padrão de 90 horas. Para testar a afirmativa dos vendedores com relação à marca **B**, foram testadas 100 de suas lâmpadas, compradas de varejistas. Esta amostra forneceu os valores $\bar{x} = 1\,140$ e $s = 80$. Como o tempo médio de vida é uma boa medida de qualidade, o problema agora é testar a hipótese de que o valor da média da marca **B** seja igual ao da marca **A** contra a hipótese alternativa de que seja um valor menor. Se a média da marca **B** é indicada por μ, este teste assumirá a forma da fórmula (3), com a desigualdade invertida, a saber

$$H_0: \mu = 1180$$
$$H_1: \mu < 1180$$

Esta alternativa foi escolhida porque sentiu-se que, se a qualidade das lâmpadas de marca **B** não fosse a mesma das lâmpadas de marca **A**, então a qualidade da marca **B** seria, sem dúvida, inferior à da marca **A**. Não é provável que os vendedores subestimem seus próprios produtos. Portanto, se estes vendedores estão falando a verdade, H_0 será verdadeira. Se eles não estão falando a verdade, sua marca será de qualidade inferior, pois nenhum vendedor seria tão bobo a ponto de afirmar apenas igualdade para seu produto, podendo realmente afirmar a superioridade dele.

Ora, o bom senso sugeriria que quanto mais à esquerda $\bar{x}$ estiver do valor médio postulado de 1 180, tanto menos fé se deverá ter na veracidade de H_0 e tanto mais fé em algum valor menor de μ como sendo a média verdadeira. Assim, é claro que a região crítica deve consistir em pequenos valores de $\bar{x}$ e por conseguinte dessa parte do eixo dos $\bar{x}$ à esquerda de algum ponto $\bar{x}_0$. O problema, conseqüentemente, é determinar o ponto $\bar{x}_0$ de forma que o valor de α seja 0,05. A técnica para fazer isso é a mesma empregada na resolução do problema do arqueólogo.

Como $n = 100$ aqui, segue-se que

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10}$$

Supõe-se que a variabilidade das lâmpadas de marca **B** seja a mesma que a de marca **A**; então $\sigma = 90$ e

$$\sigma_{\bar{x}} = \frac{90}{10} = 9$$

Desde que a área de 5 por cento da cauda esquerda de uma curva normal-padrão fique à esquerda do ponto $z = -1{,}64$, segue-se que $\bar{x}_0$ é um ponto de 1,64 desvios-padrão à esquerda da média $\mu = 1\,180$. Como o desvio-padrão aqui é $\sigma_{\bar{x}} = 9$ a desejada região crítica é essa parte do eixo dos $\bar{x}$ à esquerda de

$$1180 - 1{,}64(9) \doteq 1165$$

Estes resultados aparecem na Figura 3. Se o leitor preferir usar a álgebra para obter o valor de $\bar{x}_0$ deveria proceder como no problema do arqueólogo escrevendo a equação

$$-1{,}64 = \frac{\bar{x} - 1180}{9}$$

A resolução quanto a $\bar{x}$ dará naturalmente o resultado seguinte $\bar{x}_0 = 1\,165$.

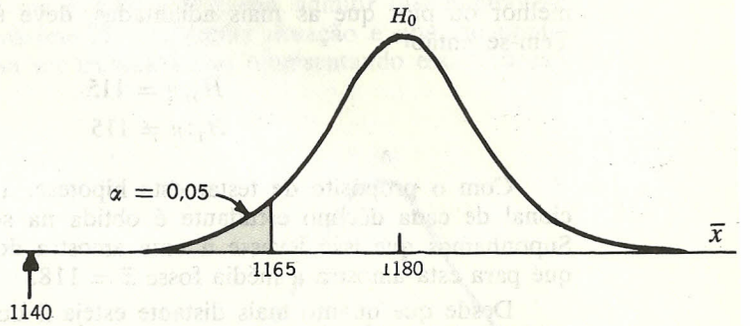


FIGURA 3. Região crítica para teste de H_0.

Agora que a região crítica foi escolhida, pode-se proceder ao teste da hipótese H_0. Desde que o valor da amostra $\bar{x} = 1\,140$ caia na região crítica, a hipótese H_0 será rejeitada. Parece muito certo que uma média da amostra tão baixa quanto 1 140 não poderia ter sido obtida de uma amostra aleatória de tamanho 100 de uma população com média 1 180. Implica isso que os vendedores das âmpadas da marca **B** não estão corretos em sua afirmativa de qualidade igual à da marca **A**. Como é certo que μ é menor que 1 180, deve-se considerar em seguida a questão em que medida ela seria menor. Se fosse desejada uma estimativa por pontos de μ, então certamente $\bar{x} = 1\,140$ seria selecionada como a estimativa. Poder-se-ia também encontrar um intervalo de confiança para μ e então determinar a máxima e a mínima diferença provável que possa haver entre as duas médias de populações. Tais considerações seriam necessárias antes que se pudesse decidir se o preço inferior para a marca **B** compensaria a qualidade inferior. Já que o objetivo desta seção é explicar como testar hipóteses, esses assuntos práticos não serão discutidos aqui; contudo, as soluções dos problemas reais pelos métodos estatísticos geralmente requerem tais considerações.

Exemplo 2. Como uma ilustração do caso em que a fórmula (2) seria preferível à (3), consideremos o problema seguinte. Registros para os vários últimos anos de calouros admitidos num certo

colégio atestam que sua nota média num teste vocacional foi de 115 e o desvio-padrão de suas notas foi de 20. Um membro da administração do colégio interessado em saber se a nova classe de calouros é típica com respeito à vocação propõe-se testar a hipótese de que a média desta nova classe seja a mesma das classes precedentes. Desde que ele não tenha razão para crer que a nova classe seja algo melhor ou pior que as mais adiantadas, deve usar a fórmula (2). Tem-se então

$$H_0: \mu = 115$$
$$H_1: \mu \neq 115$$

Com o propósito de testar esta hipótese, a nota do teste vocacional de cada décimo estudante é obtida na secretaria do colégio. Suponhamos que isso levasse a uma amostra do tamanho $n = 50$ e que para esta amostra a média fosse $\bar{x} = 118$.

Desde que quanto mais distante esteja $\bar{x}$ do valor hipotético da média 115 tanto para a direita como para a esquerda tanto menos fé se deverá ter na veracidade de H_o, é claro que a região crítica aqui deveria consistir nos valores de $\bar{x}$ em ambas as caudas da curva de $\bar{x}$ centrada em 115. Para uma amostra tão grande quanto 50, $\bar{x}$ pode ser admitida como normalmente distribuída. Além disso, se se supõe que a variabilidade das notas para os primeiros é a mesma que para os segundos, então $\sigma = 20$ e o desvio-padrão de $\bar{x}$ é dado por

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{50}} = \frac{20}{\sqrt{50}} \doteq 2{,}8$$

Desde que a probabilidade é de 0,05 de que $\bar{x}$ vá assumir um valor maior que 1,96 desvios-padrão a partir da média, segue-se que a desejada região crítica de tamanho $\alpha = 0{,}05$ deverá consistir nos valores de $\bar{x}$ nas duas caudas da curva de $\bar{x}$, exteriores aos dois valores $115 - 1{,}96(2{,}8) = 109{,}5$ e $115 + 1{,}96(2{,}8) = 120{,}5$. Estes resultados estão representados na figura 4.

Como $\bar{x} = 118$ fornece um ponto, indicado na Figura 4 por uma flecha, que não cai na região crítica, a hipótese H_o será aceita. A administração do colégio pode descansar na certeza de que a nova classe de calouros tem mais ou menos o mesmo nível de aptidão que as outras classes.

A aceitação de uma hipótese desta maneira é um caso de decisão prática. Não implica acreditar-se que a hipótese seja precisamente correta, e certamente não é uma prova da veracidade da hipótese. Antes, implica que os dados amostrais são compatíveis com

o valor postulado da média. Do ponto de vista prático, faz pouca diferença se a verdadeira média tem o valor postulado ou se tem um valor próximo ao valor postulado. Pode-se determinar quão próximo o verdadeiro valor da média deve estar do valor postulado a fim de que seja aceita a hipótese, pelos métodos do intervalo de confiança explicados no Capítulo 6. Em vista destas observações, aceitar uma hipótese vai ser interpretado como admitir que a hipótese seja razoavelmente próxima da verdadeira situação e que, do ponto de vista prático, possa ser tratada como representando essa situação verdadeira.

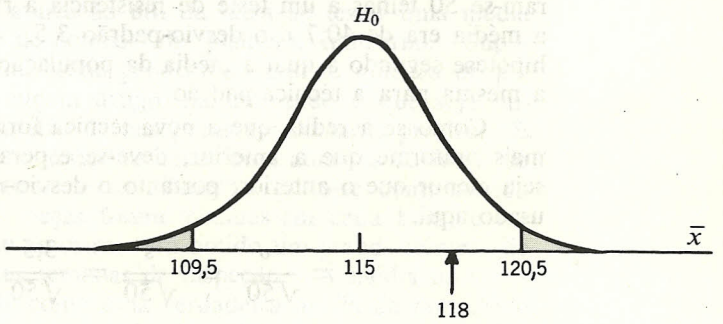


FIGURA 4. Região crítica bilateral de tamanho $\alpha = 0{,}05$.

Após o membro da administração notar que a média da amostra é maior que a antiga média, sem dúvida desejará afirmar que a nova classe é melhor que as anteriores. Em vista deste fato, haverá a tentação de tratar este problema da mesma forma como testar H_o: $\mu = 115$ em vez de H_1: $\mu > 115$ e usar uma região crítica unilateral como nos problemas precedentes; contudo, isso seria incorreto porque a decisão quanto aos valores alternativos possíveis deve ser baseada em outro conhecimento que não o fornecido pela amostra. Um modo simples de decidir se se deve usar um teste unilateral ou bilateral é perguntar quais são os valores alternativos de interesse, antes de ter sido tomada a amostra ou, o que é equivalente, antes de os resultados da amostra terem sido observados.

Em cada um dos exemplos anteriores supôs-se que o valor de σ necessário para efetuar o teste estava disponível da experiência anterior. Se tal valor não for disponível ou se houver razão para crer que a variabilidade dos valores amostrais não é a mesma que a variabilidade para primeiras amostras, então é necessário substituir σ

por sua estimativa amostral s, nos testes anteriores. Esta aproximação não leva a nenhum erro sério se a amostra for grande. Para amostras pequenas, requer-se uma técnica diferente. Será explicada na seção 6. O exemplo seguinte é um exemplo para o qual o valor de s deveria ser usado se se dispusesse de σ.

Exemplo 3. Está-se propondo uma nova técnica de produção de telhas de cerâmica apoiada no fato de que ela produz telhas de qualidade mais uniforme sem sacrificar a resistência. O processo-padrão produz telhas com uma resistência média à ruptura de 42 libras. Para testar os anúncios feitos para a nova técnica, sujeitaram-se 50 telhas a um teste de resistência à ruptura. Achou-se que a média era de 40,7 e o desvio-padrão 3,5. O problema é testar a hipótese segundo a qual a média da população para a nova técnica é a mesma para a técnica-padrão.

Como se acredita que a nova técnica forneça telha de qualidade mais uniforme que a anterior, deve-se esperar que o desvio-padrão seja menor que o anterior; portanto o desvio-padrão da amostra será usado aqui.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{50}} \cong \frac{s}{\sqrt{50}} = \frac{3{,}5}{\sqrt{50}} = 0{,}50$$

Como haveria ceticismo no tocante à afirmação de que a resistência não seria sacrificada, a região crítica deveria consistir naqueles valores de $\bar{x}$ à esquerda de

$$\mu - 1{,}64\sigma_{\bar{x}} = 42 - 1{,}64\ (0{,}50) = 41{,}2$$

Como $40{,}7 < 41{,}2$, a hipótese deve ser rejeitada. A nova técnica parece ter baixado levemente a resistência média à ruptura. Para determinar se é superior com respeito à uniformidade de qualidade, seria preciso comparar o desvio-padrão da amostra com o do processo-padrão. Esse tipo de problema é considerado num capítulo posterior.

Nas ilustrações anteriores, cada decisão para aceitar ou rejeitar H_0 foi tomada considerando-se a geometria do problema. Uma solução algébrica equivalente pode ser obtida calculando-se o valor de $z = (\bar{x} - \mu)/\sigma_{\bar{x}}$ e observando-se se é ou não compatível com H_0. Por exemplo, no problema dos crânios, $z = (194\text{-}190)/2{,}31 = 1{,}73$. Como se escolheu a região crítica como sendo os valores de $\bar{x}$, na área da cauda direita de 5 por cento de sua distribuição normal, que corresponde a valores de z maiores que 1,64 e como $1{,}73 > 1{,}64$, segue-se que H_0 deve ser rejeitada. Para o problema do exemplo 1, o valor de $z = (1140\text{-}1180)/9 = -4{,}44$. Aqui a região crítica

escolhida corresponde a valores de z menores do que $-1{,}64$; conseqüentemente, H_0 deve ser rejeitado porque $-4{,}44 < -1{,}64$. No exemplo 2, $z\ (118\text{-}115)/2{,}8 = 1{,}07$. Como se escolheu uma região crítica bicaudal os valores críticos correspondentes de z consistem em $|z| > 1{,}96$; portanto deve-se aceitar H_0.

As respostas no apêndice são expressas nesta forma algébrica. Dão o valor de z e depois a decisão adequada, baseada neste valor.

Os tamanhos dos erros do tipo II nestes problemas não foram calculados porque exigiriam muito mais cálculo e discussão de problemas.

Exemplo 4. Uma aplicação útil da idéia de testar uma média apresenta-se no trabalho de controle de qualidade industrial. Suponhamos que uma máquina esteja produzindo grande número de peças que são usadas em algum artigo manufaturado e que seja importante o fato de o diâmetro de tais peças ser bastante preciso. É costume que um fiscal tire periodicamente uma amostra da linha de produção a fim de verificar se os diâmetros estão-se comportando convenientemente. Se 5 peças forem medidas em cada hora, e for registrada a sua média de amostra, será obtido um grande número de valores de $\bar{x}$ após algumas semanas de inspeção. A média de todos esses $\bar{x}$ poderá ser tratada como uma verdadeira média da população dos diâmetros das peças e o seu desvio-padrão como o verdadeiro valor de $\sigma_{\bar{x}}$.

Pelo tratamento dos valores precedentes como os verdadeiros valores, pode-se calcular os valores $\mu - 3\sigma_{\bar{x}}$ e $\mu + 3\sigma_{\bar{x}}$. Das probabilidades da curva normal, a probabilidade é de 0,997 de que um valor de $\bar{x}$ da amostra vá recair entre esses dois limites; por conseguinte se um valor de $\bar{x}$ cai fora deste intervalo, há boa razão para acreditar que algo andou errado com a máquina ao fabricar as peças. A experiência tem demonstrado que uma máquina operando convenientemente terá um comportamento muito semelhante ao de uma máquina de número aleatório no sentido de que sucessivas peças produzidas são como que amostras aleatórias de uma população de peças. Usa-se um intervalo de três desvios-padrão porque cerca de 5 por cento dos valores de $\bar{x}$ cairão fora deste último intervalo, mesmo que a máquina trabalhe satisfatoriamente, e conseqüentemente estaria procurando dificuldade sem razão para isso. Além disso, uma máquina industrial é apenas uma aproximação de uma máquina ideal de números aleatórios. A experiência indica que um intervalo de três desvios-padrão é razoavelmente justo, do ponto de vista prático.

Após os dados terem sido reunidos por algumas semanas de forma que os limites $\mu - 3\sigma_{\bar{x}}$ e $\mu + 3\sigma_{\bar{x}}$ possam ser obtidos, o diagrama de controle está pronto para ser construído. Consiste simples-

mente em uma faixa horizontal com os números aleatórios marcados no eixo horizontal. Um gráfico de controle desse tipo é visto na Figura 5.

Cada ponto corresponde a um valor $\bar{x}$ de amostra obtido de 5 peças selecionadas cada hora, após a coleta inicial dos dados. Será observado que o processo parece estar sob controle. A vantagem gritante de um gráfico de controle é que ele adverte o inspetor por meio de probabilidade das dificuldades com uma máquina antes de ter produzido um número grande de peças defeituosas, o que de outra forma só poderia ser descoberto algum tempo depois quando as peças já começassem a ser usadas na constituição do artigo manufaturado.

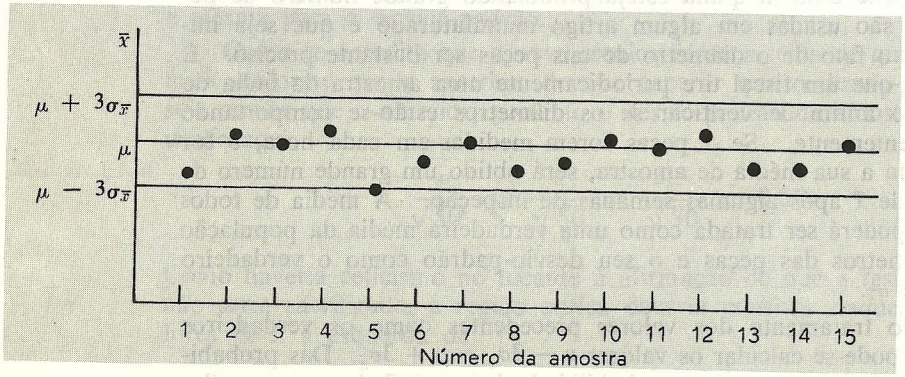


FIGURA 5 Gráfico de controle da média.

3 — TESTE DE PROPORÇÃO

Os métodos da curva normal de grandes amostras empregados para resolver problemas de estimação para o binomial p podem ser empregados também para testar hipóteses sobre p. Como resultado, a técnica para testar as hipóteses pelas quais p tem um valor fixo é exatamente a mesma que a explicada na última seção para testar μ.

Exemplo 1. Como ilustração, consideremos o seguinte problema de genética. Conforme a teoria da herança mendeliana, certos cruzamentos de ervilhas deveriam dar ervilhas amarelas e verdes numa razão de 3:1. Num experimento foram obtidas 176 ervilhas amarelas e 48 ervilhas verdes. Esses números são compatíveis com a teoria mendeliana?

Este problema pode ser considerado como o problema de testar a hipótese

$$H_0: p = \frac{3}{4}$$

Onde p *significa* a probabilidade de que uma ervilha escolhida ao acaso seja amarela. As 224 ervilhas podem ser tratadas como 224 tentativas de um experimento para o qual $p = \frac{3}{4}$ é a probabilidade de sucesso de uma única tentativa. Da fórmula (7), Capítulo 5, segue-se que $\hat{p} = x/n$ pode ser tratada como variável normal com média $p = \frac{3}{4}$ e desvio-padrão dado por

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{3}{4} \cdot \frac{1}{4}}{224}} = 0{,}029$$

O problema agora é muito semelhante ao problema do teste da média normal. A região crítica aqui é escolhida como aqueles valores de $\hat{p}$ nas duas caudas da curva normal de $\hat{p}$. Para $\alpha = 0{,}05$, a região crítica consistirá dos valores de $\hat{p}$ que caem fora do intervalo dado por

$$p - 1{,}96\sqrt{\frac{pq}{n}} \quad \text{e} \quad p + 1{,}96\sqrt{\frac{pq}{n}}$$

Como $n = 224$ e $p = \frac{3}{4}$ aqui, os cálculos darão o intervalo de (0,693, 0,807). A Figura 6 mostra a distribuição normal aproxi-

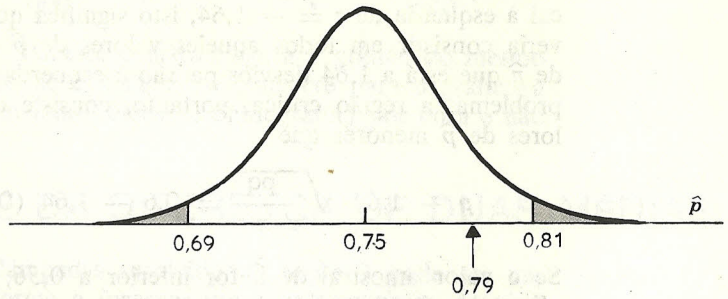


FIGURA 6. Distribuição normal aproximada de $\hat{p}$.

mada para $\hat{p}$ e a região crítica determinada. Como o valor da amostra $\hat{p} = 176/224 = 0{,}79$ não cai na região crítica, aceita-se a hipótese H_0. Assim, com base nesses dados, não há razão para se duvidar que a herança mendeliana esteja operando aqui.

Exemplo 2. Como segunda ilustração para usar métodos de curva normal com o fim de testar uma hipótese sobre p binomial, consideremos o problema discutido brevemente nos fins do Capítulo 5.

Certo político afirmou ter um apoio de 60 por cento com relação a uma proposta legislativa e colheu-se então uma amostra de 400 votantes para testar sua afirmativa. Surgiu então o problema sobre qual deveria ser o tamanho percentual da amostra antes de se refutar mediante raciocínio a afirmativa. A questão pode ser considerada como um problema de testar a hipótese

$$H_0 : p = 0{,}6$$

Uma vez que o interesse neste problema centra-se no fato de $p = 0{,}6$, contra a possibilidade de $p < 0{,}6$, este problema é um pouco como o das lâmpadas, discutido na primeira seção deste capítulo, no qual as alternativas estão todas de um lado do valor hipotético e portanto a região crítica deveria estar sob uma das caudas da respectiva curva normal. A hipótese alternativa natural aqui é

$$H_1 : p < 0{,}6$$

A região crítica de tamanho $\alpha = 0{,}05$ para este problema devia portanto ser selecionada para estar sob os 5 por cento da cauda esquerda da curva normal cuja média é $p = 0{,}6$ e cujo desvio-padrão é dado por:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0{,}6)\ (0{,}4)}{400}} = 0{,}0245$$

Como, da Tabela IV, 5 por cento da área de uma curva normal-padrão cai à esquerda de $z = -1{,}64$, isto significa que a região crítica deveria consistir em todos aqueles valores de $\hat{p}$ menores que o valor de $\hat{p}$ que está a 1,64 desvios-padrão à esquerda da média. Para este problema, a região crítica, portanto, consiste em todos aqueles valores de $\hat{p}$ menores que

$$p - 1{,}64\sqrt{\frac{pq}{n}} = 0{,}6 - 1{,}64\ (0{,}245) = 0{,}56$$

Se o valor amostral de $\hat{p}$ for inferior a 0,56, dever-se-ia rejeitar a afirmativa do político.

Outra aplicação útil do teste do p binomial surge em gráficos de controle industrial para a percentagem de peças defeituosas, em produção de peças em massa. Esta aplicação não se destina a problemas industriais. Pode-se usar sempre que se tenham operações repetidas. A técnica é exatamente a mesma que a de gráficos de controle para a média. Usa-se experiência acumulada para se obter uma boa estimativa de p; constrói-se depois a faixa de controle dada por $p - 3\sqrt{pq/n}$ e $p + 3\sqrt{pq/n}$. Aqui n é a amostra do tamanho sobre o qual está baseada cada proporção do gráfico.

Exemplo 3. Para exemplificar como se deveria construir tal gráfico, vamos considerar o seguinte problema. Registra-se durante 10 dias o número de erros de datilografia cometidos por estudantes que estão aprendendo a escrever a máquina. Durante este período bateram 20 000 palavras, das quais 800 estavam erradas. O problema consiste em usar estes dados para construir um gráfico de controle para a proporção de erros cometidos, por hora, por um estudante que bate aproximadamente 600 palavras por hora. Supondo que o mencionado estudante seja típico, temos uma boa estimativa de p dividindo o número total de palavras erradas pelo número total de palavras datilografadas. Esta estimativa é

$$p = \frac{800}{20\,000} = 0{,}04$$

Como as proporções a serem colocadas no gráfico de controle são as que surgem num período de datilografia de uma hora, segue-se que $n = 600$ aqui. Se estes valores forem substituídos nas fórmulas apresentadas no parágrafo anterior, os limites inferior e superior do gráfico de controle serão

$$0{,}04 - 3\sqrt{\frac{(0{,}04)\ (0{,}96)}{600}} \text{ e } 0{,}04 + 3\sqrt{\frac{(0{,}04)\ (0{,}96)}{600}}$$

isso dá 0,016 e 0,064. Pode-se construir agora o gráfico do mesmo modo que para a média, só que agora se transfere para o gráfico a proporção de palavras erradas (erros de datilografia) por hora e não a média da amostra.

4 — TESTE DA DIFERENÇA DE DUAS MÉDIAS

O problema das lâmpadas resolvido na seção 2 pode ser um pouco modificado de forma a fornecer um problema típico da vida real. Suponhamos que a cidade que estivesse comprando lâmpadas não tivesse experiência nem com a marca A nem com a marca B e

desejasse decidir qual a marca a comprar, sendo os preços os mesmos. Seria então necessário testar uma amostra de cada marca ao invés de testar justamente uma amostra da marca B como no problema anterior. Suponhamos que seja testada uma amostra de 100 lâmpadas de cada uma das duas marcas e que estas amostras dêem os valores $\bar{x}_1 = 1\,160$, $s_1 = 90$, $\bar{x}_2 = 1\,140$ e $s_2 = 80$ onde os índices 1 e 2 referem-se às marcas A e B respectivamente.

Agora se a marca A dá uma vida média maior que a B, parecerá que ela é superior; contudo poderia ser verdade o inverso devido, entretanto, a um azar com algumas das lâmpadas da marca B por ter sido produzida uma média de amostra desusadamente baixa. Poderia suceder que um segundo conjunto de 100 amostras de cada marca fosse produzir resultados bem diferentes. O problema conseqüentemente se reduz a determinar se esta diferença de médias de amostra, a saber, $\bar{x}_1 - \bar{x}_2$, é grande o bastante para justificar a crença de que a marca A seja superior à B.

A fim de resolver este problema, é necessário saber como $\bar{x}_1 + - \bar{x}_2$ varia ao serem realizados repetidos experimentos de amostragem da mesma espécie. Cada um desses experimentos consiste em tomar uma amostra de 100 lâmpadas das marcas A e B, testar as 200 lâmpadas, determinar os valores de $\bar{x}_1$ e $\bar{x}_2$ e registrar o valor da variável $\bar{x}_1 - \bar{x}_2$. Se fosse levado a efeito um grande número de tais experimentos de amostragem seria obtido um grande número de valores da variável $\bar{x}_1 - \bar{x}_2$. Estes valores poderiam ser classificados numa tabela de freqüência, sendo traçado o histograma para dar uma boa idéia da distribuição-limite ou teórica de $\bar{x}_1 - \bar{x}_2$. Como no caso de uma média isolada, $\bar{x}$, não é necessário realizar estes experimentos de amostragem, porque a forma de distribuição-limite pode ser elaborada matematicamente. Pode-se verificar que $\bar{x}_1 - \bar{x}_2$ possuirá uma distribuição normal se x_1 e x_2 a possuírem com uma média igual à diferença das médias das populações de x_1 e x_2, a saber, $\mu_1 - \mu_2$, e com variância (quadrado do desvio-padrão) igual à soma das variâncias $\bar{x}_1$ e $\bar{x}_2$. Este resultado é expresso na forma de um teorema:

(4) **Teorema.** *Se* x_1 *e* x_2 *possuem distribuições normais independentes com médias* μ_1 *e* μ_2, *e desvios-padrões* σ_1 *e* σ_2 *então a variável* $\bar{x}_1 - \bar{x}_2$ *possuirá uma distribuição normal com média* $\mu_1 - \mu_2$ *e desvio-padrão dado pela fórmula*

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Aqui n_1 e n_2 são os tamanhos de amostra nas quais $\bar{x}_1$ e $\bar{x}_2$ são baseadas. Deve-se notar que o teorema requer que as duas populações donde estão sendo tiradas as amostras sejam normais; contudo, se n_1 e n_2 são, digamos, de tamanho 25, então $\bar{x}_1$ e $\bar{x}_2$ podem ser admitidas como normalmente distribuídas, sucedendo então o mesmo com $\bar{x}_1 - \bar{x}_2$. Deve-se ainda notar que é exigido que x_1 e x_2 sejam variáveis independentes. Por exemplo, se o problema tivesse sido testar a hipótese de que não há diferença entre o comprimento médio dos pés direito e esquerdo de pessoas, não seria correto deixar x_1 representar o comprimento do pé direito de um indivíduo e x_2 o comprimento de seu pé esquerdo, pois um grande valor de x_1 certamente aumentaria a probabilidade de um grande valor de x_2. Se, contudo, os valores de x_2 fossem obtidos de um conjunto diferente de indivíduos que não fosse o de x_1, então x_1 e x_2 seriam variáveis independentes. O modo conveniente de tratar dados duplos, tais como os decorrentes da medida de ambos os pés de cada indivíduo, seria tomar a diferença de cada par e usar então os métodos anteriores para testar se a média da variável $x_1 - x_2$ é zero. Esta técnica será discutida na seção 6.

Problemas como a seleção de lâmpadas podem ser resolvidos por meio do Teorema (4), contanto que sejam tratados como problemas de testar a hipótese adequada. De um ponto de vista prático, deveria haver pouca diferença se rejeitássemos a hipótese de que as marcas sejam igualmente boas ou a hipótese de que elas difiram em qualidade. Como resultado, estabelece-se a hipótese

(5) $$H_0: \mu_1 = \mu_2$$

Um modo equivalente de escrevermos a equação é:

$$H_0: \mu_1 - \mu_2 = 0$$

Este tipo de hipótese é conhecido como hipótese *nula,* porque supõe que não haja diferença. Muito freqüentemente, contudo, o experimentador acredita que há uma diferença apreciável e espera que a evidência da amostra vá rejeitar a hipótese. Se a amostra rejeitar a hipótese, então pode-se afirmar justificadamente que existe uma diferença real nas médias das populações. Se a amostra não rejeitar a hipótese, então pode-se afirmar justificadamente que existe uma diferença real nas médias das populações. Se a amostra não rejeitar a hipótese, então há uma probabilidade de que a diferença da amostra seja causada pela variação na amostragem, admitida a suposição de que as médias das populações sejam iguais.

À luz do Teorema (4) e da discussão precedente, o teste da hipótese dada pela fórmula (5) é equivalente ao teste da hipótese de

que a média da variável normal $\bar{x}_1 - \bar{x}_2$ seja zero. Mas este tipo de problema foi resolvido na seção 2. Uma vez que a hipótese alternativa ordinariamente seria escolhida como:

$$H_1: \mu_1 \neq \mu_2$$

segue-se que se deveriam aplicar aqui os mesmos métodos da solução do problema do teste vocacional que usou a fórmula (2). Agora, do Teorema (4) a variável $\bar{x}_1 - \bar{x}_2$ pode ser admitida como normalmente distribuída como média 0, porque $\mu_1 = \mu_2$ sob H_o, e com desvio-padrão dado por

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Infelizmente os valores da população ${\sigma_1}^1$ e ${\sigma_2}^2$ são desconhecidos; conseqüentemente, devem ser aproximados por suas estimativas de amostra, a saber $s_1^2 = (90)^2$ e $s_2^2 = (80)^2$. Como $n_1 = n_2 = 100$ aqui, o valor aproximado do desvio-padrão torna-se

$$\sigma_{\bar{x}_1 - \bar{x}_2} \doteq \sqrt{\frac{8100}{100} + \frac{6400}{100}} \doteq 12$$

Escolhe-se uma região crítica, baseada em áreas de caudas iguais sob a curva normal para $\bar{x}_1 - \bar{x}_2$, para a qual $\alpha = 0{,}05$. O que significa que a região crítica consiste naquela parte do eixo horizontal que cai mais de 1,96 $\sigma_{\bar{x}_1 - \bar{x}_2} = 24$ unidades afastada de 0. A Figura 7 mostra geometricamente a distribuição de $\bar{x}_1 - \bar{x}_2$ e a região crítica escolhida.

Uma vez que as duas amostras de 100 cada uma deram o valor $\bar{x}_1 - \bar{x}_2 = 1\,160 - 1\,140 = 20$ e como 20 cai na região crítica para este teste, a hipótese será aceita.

Justamente como no problema anterior de testar uma média isolada, a aceitação de uma hipótese desse modo não implica que se acredite que a hipótese seja verdadeira. Implica, contudo, não se estar convencido, pela evidência da amostra, de que haja uma diferença apreciável e que, a menos que seja apresentada uma maior evidência em contrário, se deseje admitir que para todos os propósitos práticos não haja diferença apreciável nas médias de população. É uma conveniência matemática formular uma hipótese desta maneira. Seria mais realista testar se as médias diferiam por menos que uma quantidade especificada, mas a teoria resultante seria muito mais complicada. O estudante não deverá iludir-se acreditando que provou ser verdadeira a hipótese pelo simples fato de ter concordado em aceitá-la.

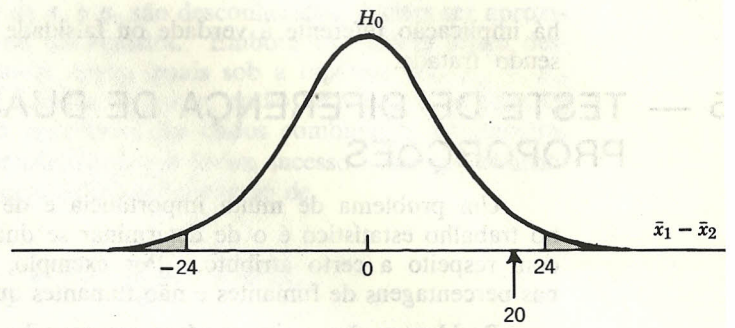


FIGURA 7. Distribuição de $\bar{x}_1 - \bar{x}_2$, quando $\mu_1 = \mu_2$.

Em vista de o fato das estimativas de amostra terem sido necessárias como aproximações para as variâncias de população, os métodos aqui usados são métodos de grande amostra. Um método destinado à solução de problemas de pequenas amostras será discutido na seção 6.

Há várias palavras e frases usadas em ligação com o teste de hipóteses para as quais devemos chamar a atenção dos estudantes. Quando um teste de hipótese produz um valor da amostra que cai na região crítica do teste, diz-se que o resultado é *significativo;* caso contrário, o resultado é *não-significativo.*

Esta palavra surge do fato de o valor da amostra que cai na região crítica não ser compatível com a hipótese, e portanto significar a necessidade de alguma outra hipótese. A probabilidade de se cometer um erro do tipo I, simbolizado por α, chama-se *nível de significância* do teste. Para problemas resolvidos rotineiramente neste livro, o nível de significância foi escolhido como sendo igual a 0,05.

Se analisarmos a técnica usada para testar as várias hipóteses tratadas assim, observaremos que é uma simples técnica de tomada de decisão. Esta regra se baseia no valor amostral de alguma variável aleatória e consiste em dividir todos os valores possíveis da variável aleatória em dois grupos, os associados à rejeição de H_0 e que formam o que se chama de região crítica do teste e os associados à aceitação de H_0. Deste ponto de vista geral, um teste de hipótese é uma maneira meramente sistemática de se tomar uma decisão prática e não

há implicação referente à verdade ou falsidade da hipótese que está sendo tratada.

5 — TESTE DE DIFERENÇA DE DUAS PROPORÇÕES

Um problema de muita importância e de ocorrência freqüente no trabalho estatístico é o de determinar se duas populações diferem com respeito a certo atributo. Por exemplo, há alguma diferença nas percentagens de fumantes e não fumantes que sofrem do coração?

Problemas desse tipo podem ser tratados como problemas de testar a hipótese

$$H_0: p_1 = p_2$$

na qual p_1 e p_2 são as duas proporções do atributo da população. Se n_1 e n_2 indicam o tamanho das amostras e $\hat{p}_1$ e $\hat{p}_2$ as proporções resultantes da amostra, então a variável a usar na resolução desse problema é $\hat{p}_1 - \hat{p}_2$. Isso corresponde a usar $\bar{x}_1 - \bar{x}_2$ no problema de testar a hipótese de que $\mu_1 = \mu_2$. Os métodos usados para resolver esse problema podem ser empregados aqui também porque $\hat{p}_1 - \hat{p}_2$ pode ser tratado como duas variáveis normais independentes, com média $p_1 - p_2$ e desvio-padrão dado por

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Ao testar a hipótese $H_0 : p_1 = p_2$, a média da distribuição de $\hat{p}_1 - \hat{p}_2$ será, naturalmente, igual a 0.

Exemplo 1. Para ilustrar como usar essas fórmulas, consideremos o seguinte problema. Uma amostra de 400 marinheiros foi dividida em dois grupos iguais por seleção aleatória. A um grupo foram dadas pílulas da marca **A** de um preventivo contra enjôo e ao outro grupo, pílulas da marca **B**. O número de cada grupo que não enjoou durante uma forte tempestade foi de 152 e 132. Pode-se concluir que não há diferença real na eficácia desses pílulas?

Os cálculos dão

$$\hat{p}_1 = \frac{152}{200} = 0{,}76 \qquad \hat{p}_2 = \frac{132}{200} = 0{,}66$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{200} + \frac{p_2 q_2}{200}}$$

Como os valores de p_1 e p_2 são desconhecidos, podem ser aproximados pelas estimativas da amostra. Embora os valores sejam desconhecidos, são admitidos como iguais sob a hipótese H_0: $p_1 = p_2$. Se esse valor comum é indicado por p, então uma boa estimativa de p é o valor obtido da proporção dos dados combinados da amostra. Houve 284 dos 400 marinheiros que foram sucesso neste experimento total; dessa forma p seria estimado por meio de

$$\hat{p} = \frac{284}{400} = 0{,}71$$

Substituindo-se p_1 e p_2 por p na fórmula para o desvio-padrão e depois aproximando-se p de p, obtemos

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{(0{,}71)\ (0{,}29)\ \left(\frac{1}{200} + \frac{1}{200}\right)} = 0{,}045$$

A adoção de uma região crítica bilateral de tamanho 0,05 conduz à região crítica representada na Figura 8. Como $\hat{p}_1 - \hat{p}_2 = 0{,}10$, aqui, cai na região crítica, a hipótese é rejeitada. Parece que a marca **A** oferece uma proteção um tanto melhor contra o enjôo do que a marca **B**, pelo menos para marinheiros em tempestade.

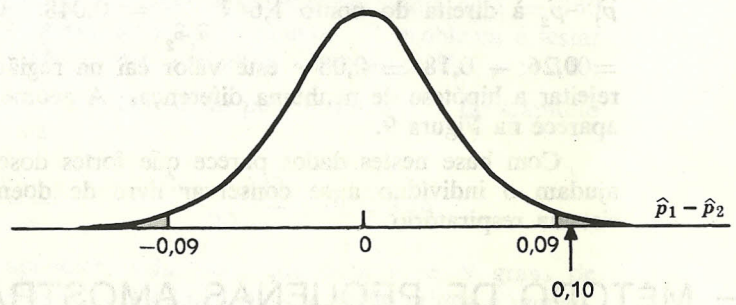


FIGURA 8. Distribuição e região crítica para $\hat{p}_1 - \hat{p}_2$ no problema do enjôo

Exemplo 2. Um artigo do **Canadian Medical Association Journal,** de novembro de 1972, discute os efeitos de fortes doses de vitamina C sobre os resfriados. No experimento efetuado, houve 407 indivíduos que tomaram pesadas doses de vitamina C e 411 que

tomaram pseudomedicamentos só para efeitos psicológicos. No grupo da vitamina, 105 dos participantes ficaram livres de doenças de tipo respiratório durante o tempo da análise, ao passo que no grupo dos pseudomedicamentos só 76 o ficaram. Teste a hipótese segundo a qual não há diferença nas verdadeiras proporções dos dois grupos.

Estes dados fornecem os valores

$$\hat{p}_1 = \frac{105}{407} = 0{,}26 \quad \text{e} \quad \hat{p}_2 = \frac{76}{411} = 0{,}18$$

Como a hipótese é $H_0 : p_1 = p_2$, uma estimativa de $p = p_1 = p_2$ é obtida pela combinação dos dois resultados amostrais. Isso dá

$$\hat{p} = \frac{181}{818} = 0{,}22$$

Dessa forma, uma estimativa de $\sigma_{\hat{p}_1 - \hat{p}_2}$ é dada por

$$\sigma_{\hat{p}_1 - \hat{p}_2} \cong \sqrt{(0{,}22)\,(0{,}78)\left[\frac{1}{407} + \frac{1}{411}\right]} \cong 0{,}029$$

Como a hipótese alternativa aqui é $H_1 : p_1 > p_2$, a cauda direita da distribuição de $\hat{p}_1 - \hat{p}_2$ deve ser escolhida como região crítica para este teste. Isto significa que a região crítica deve consistir nos valores $\hat{p}_1 - \hat{p}_2$ à direita do ponto $1{,}64\,\sigma_{\hat{p}_1 - \hat{p}_2} = 0{,}048$. Como $\hat{p}_1 - \hat{p}_2 = 0{,}26 - 0{,}18 = 0{,}08$ e este valor cai na região crítica, deve-se rejeitar a hipótese de nenhuma diferença. A geometria do problema aparece na Figura 9.

Com base nestes dados parece que fortes doses de vitamina C ajudam o indivíduo a se conservar livre de doenças relativas ao sistema respiratório.

6 — MÉTODO DE PEQUENAS AMOSTRAS

Nos exemplos da seção 2 tínhamos o valor de σ da experiência antiga. Como foi dito então, se a experiência anterior não oferece o valor de σ, é necessário substituir σ pela estimativa s de sua amostra. Isto é válido para grandes amostras, mas pode surgir um erro grave desta aproximação se n for pequeno. Para tais problemas de pequenas amostras, pode-se usar a distribuição t de Student da mesma forma que se usa o z normal para grandes amostras.

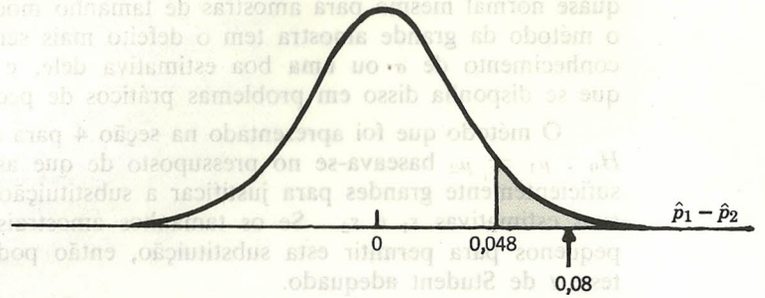


FIGURA 9. Distribuição e região crítica para $\hat{p}_1 - \hat{p}_2$ no problema da vitamina C.

Exemplo 1. Como ilustração de um problema de amostra pequena, consideremos o seguinte exemplo. Está sendo testada uma nova fórmula de um combustível para mísseis a fim de se determinar se é ou não superior à fórmula corrente. De experiências passadas, a distância média percorrida por mísseis experimentais lançados ao Oceano Pacífico era de 340 milhas. Como se espera que o novo combustível afete a variabilidade da distância percorrida e também a distância média, o valor de σ baseado em experiência passada não será usado aqui. Dez mísseis experimentais com o novo combustível foram lançados no Pacífico e produziram uma distância média de $\bar{x} = 360$ milhas e um desvio-padrão de $s = 20$ milhas. O problema é testar a hipótese $H_0 : \mu = 340$ contra a hipótese alternativa $H_1 : \mu > 340$.

Basta calcular o valor de t dado pela fórmula (7) do Capítulo 7. Para este problema

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{360 - 340}{20}\sqrt{10} = 3{,}16$$

Da Tabela V no apêndice, achar-se-á que para $\nu = 9$ graus de liberdade o valor 0,05 de t é 1,833. Como a região crítica é unicaudal e consiste em valores de $t > 1{,}833$, o valor de $t = 3{,}16$ cai na região crítica, e portanto rejeita-se H_0. O novo combustível aumentou, sem dúvida, a distância média percorrida.

Como a derivação t de Student requer a suposição de que a variável básica x esteja normalmente distribuída, ao aplicarmos isso a pequenas amostras devemo-nos assegurar de que x possui uma distribuição normal aproximada. O método da grande amostra não exige tal precaução, porque provavelmente $\bar{x}$ tem uma distribuição

quase normal mesmo para amostras de tamanho moderado; contudo, o método da grande amostra tem o defeito mais sério de exigir um conhecimento de σ ou uma boa estimativa dele, e não é provável que se disponha disso em problemas práticos de pequenas amostras.

O método que foi apresentado na seção 4 para testar a hipótese $H_0 : \mu_1 = \mu_2$ baseava-se no pressuposto de que as amostras eram suficientemente grandes para justificar a substituição de σ_1 e σ_2 por suas estimativas s_1 e s_2. Se os tamanhos amostrais são demasiado pequenos para permitir esta substituição, então pode-se aplicar um teste t de Student adequado.

Para testarmos a diferença de duas médias, a teoria da distribuição t de Student exige que se suponha que duas variáveis básicas x_1 e x_2 possuam distribuições normais independentes com desvios-padrão iguais. Essas restrições são consideravelmente mais restritivas do que as necessárias para métodos de grande amostra. Se estes pressupostos forem razoavelmente satisfeitos, pode-se então tratar a variável

$$(6) \qquad t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

como uma variável t de Student com $v = n_1 + n_2 - 2$ graus de liberdade. A solução é então levada a cabo da mesma maneira que para o teste de uma única média.

Exemplo 2. Um artigo no jornal médico *Lancet*, de julho de 1973, discute uma experiência levada a efeito por dois professores de um hospital de Londres no sentido de medir a quantidade de tabaco inalada por fumantes de cachimbo e charuto. Seu experimento consistiu em medir (1) a quantidade de carboxi-hemoglobina, **COHb**, no sangue de 11 fumantes de cachimbo e charuto, mas que nunca tinham fumado cigarro, e (2) a quantidade de **COHb** no sangue de 39 pessoas que fumaram cigarro por muito tempo, mas que já tinham trocado o cigarro por cachimbo ou charuto muitos anos antes. As médias e os desvios-padrão para os dois grupos são

$$\bar{x}_1 = 2{,}3 \quad \bar{x}_2 = 5{,}2 \quad s_1 = 1{,}0 \quad s_2 = 2{,}7$$

Como $\mu_1 = \mu_2$ sob a hipótese que está sendo testada, o valor de t como na fórmula (6) é

$$t = \frac{2{,}3 - 5{,}2}{\sqrt{10(1{,}0)^2 + 38(2{,}7)^2}} \sqrt{\frac{11 \cdot 39 \cdot 48}{50}} = -3{,}5$$

Se não houver razão para crer que um grupo tenha um nível mais baixo de **COHb** que outro, deve-se usar um teste bicaudal, onde o valor crítico esquerdo de t para 48 graus de liberdade deve ser achado como sendo aproximadamente — 2,0. Se o experimentador estiver convencido antes de começar o experimento, de que os não fumantes de cigarro devem ter um nível mais baixo, deve então usar um teste unicaudal e comparar — 3,5 com o valor crítico de aproximadamente — 1,68. Em ambos os casos o resultado é significativo e parece certo que mudar para cachimbo ou charuto depois de se ter fumado cigarro, por muito tempo, não faz com que o nível de **COHb** do sangue se iguale ao dos que nunca fumaram cicarros. Sabe-se que o nível de **COHb** para os que fumam cigarros e que nunca os abandonaram é semelhante ao daqueles que mudaram para cachimbo ou charuto. A **COHb** está diretamente implicada na gênese da arteriosclerose que, por sua vez, está fortemente associada a problemas sérios do coração.

Pares Combinados

Em alguns experimentos destinados a testar se duas médias populacionais são iguais, o investigador elabora seu experimento de modo a comparar tipos semelhantes de indivíduos na esperança de fazer mais eficiente o teste. Por exemplo, se um experimento está sendo efetuado para se determinar se dois métodos diferentes de ensinar aritmética fornecem resultados diferentes, é muitas vezes aconselhável selecionar pares de estudantes que antes tiveram notas semelhantes nesta matéria e que se assemelham também sob outros aspectos. Se as notas conseguidas pelo i-ésimo par de tais estudantes forem simbolizadas por x_i e y_i e dispusermos de n de tais pares, então trata-se a variável $z_i = x_i - y_i$ como o i-ésimo valor amostral numa amostra de tamanho n de uma variável normal e testa-se a hipótese de que a medida desta variável é zero. Assim o teste se reduz a testar uma média normal única. Supõe-se aqui que após a constituição dos pares, cada um dos estudantes de cada par seja selecionado, ao acaso, para ir à classe onde se ensina pelo primeiro e pelo segundo método respectivamente. Como é provável que esta escolha de estudantes em pares elimine muito da variabilidade no desempenho dos estudantes, um teste deste tipo é mais apto a provar a existência de qualquer diferença, do que o seria o teste para o qual se destina a fórmula (6) e que requer dois conjuntos de n estudantes selecionados aleatoriamente. O teste anterior é geralmente conhecido como *teste dos pares*

ordenados. Após a obtenção das diferenças deve ser efetuado como o teste do Exemplo 1.

7 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Suponha que x possua uma distribuição normal com $\sigma = 6$ mas que sua média seja desconhecida. Se tomarmos uma amostra de tamanho 16 e $\bar{x} = 33$, (a) 2 — Teste a hipótese $H_0 : \mu = 30$ contra $H_1 : \mu > 30$. (b) 4 — Tomou-se uma segunda amostra de tamanho 32, algum tempo mais tarde e obteve-se uma média de 29; dessa forma há razão para crer que a média mudou. Teste a hipótese de que a média não mudou contra a possibilidade de ela ter-se deslocado para a esquerda. Seguem-se as soluções.

(a) $$\sigma_{\bar{x}} = \frac{6}{\sqrt{16}} = 1{,}5 \qquad z = \frac{33 - 30}{1{,}5} = 2{,}00$$

Como grandes valores de $\bar{x}$ favoreceriam H_1, a região crítica deveria consistir em grandes valores de $\bar{x}$ e dessa forma grandes valores de z. Para $\alpha = 0{,}05$ a região crítica é portanto dada por $z > 1{,}64$. Como $z = 2{,}00$ cai na região crítica, rejeita-se H_0.

(b) $$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{36\left[\frac{1}{16} + \frac{1}{32}\right]} = 1{,}84$$

$$z = \frac{29 - 33}{1{,}84} = -2{,}17$$

Aqui a cauda esquerda da distribuição de $\bar{x}_1 - \bar{x}_2$ deveria ser escolhida como região crítica porque favorece o deslocamento para a esquerda. Para $\alpha = 0{,}05$ isto equivale a escolher $z < -1{,}64$ como a região crítica. Como $z = -2{,}17$ cai na região crítica, rejeitamos H_0.

Exemplo 2. Um inquérito, relatado no *Los Angeles Times,* de 11 de julho de 1974, feito por Irvine, professor de psicologia médica de uma universidade da Califórnia, em conjunto com o *Mental Health Department* de Orange, determinou o número de motoristas bêbados entre os motoristas de fim de semana da comarca de Orange. A média nacional está próxima de 5 por cento. Entre 1 000 motoristas parados e testados em Orange, 7 por cento foram classificados como bêbados. A definição, em ambos os casos, baseava-se na quantidade de álcool no sangue, e portanto as duas percentagens deveriam ser comparáveis. Teste a hipótese de que $p = 0{,}05$ para os motoristas de Orange.

Aqui $\hat{p} = 0{,}07$ e $n = 1\,000$; dessa forma

$$z = \frac{0{,}07 - 0{,}05}{\sqrt{\frac{(0{,}05)\,(0{,}95)}{1\,000}}} = 2{,}9$$

Este valor de z excede o valor crítico de 1,96 baseado num teste bicaudal; portanto, a hipótese de que a percentagem de Orange é a mesma que a percentagem nacional, deve ser rejeitada. Há, sem dúvida, uma percentagem maior de motoristas bêbados nos "weekends" de Orange que na nação como um todo.

Exemplo 3. Uma amostra de proprietários de carro (amostra de tamanho 200) numa certa cidade mostrou que 48 deles tinham carteiras de habilitação já vencidas. A proporção de tais motoristas para o ano anterior foi de 0,30. Usando estes dados resolva os problemas seguintes: (a) 3 — Teste a hipótese de que a proporção da população é $p = 0{,}30$. (b) 3 — Encontre os limites de um gráfico de controle para $\hat{p}$ se se deve tomar amostras diárias de tamanho 50 e se a proporção da população for de $p = 0{,}30$.

(a) $\hat{p} = \frac{48}{200} = 0{,}24$; dessa forma

$$z = \frac{0{,}24 - 0{,}30}{\sqrt{\frac{(0{,}30)\,(0{,}70)}{200}}} = -1{,}85$$

Como $|z| < 1{,}96$, aceita-se H_0.

(b) Como $n = 50$, os limites são dados por

$$p \pm 3\sqrt{pq/50} = 0{,}30 \pm 3\sqrt{\frac{(0{,}30)\,(0{,}70)}{50}} = 0{,}30 \pm 0{,}19$$

dessa forma são 0,11 e 0,49

Exemplo 4. O *Newsweek* de 27 de abril de 1974 relatou um experimento levado a cabo pela *Reproductive Biology Research Foundation,* de S. Luís, para determinar os efeitos da maconha sobre a sexualidade. No experimento, selecionaram-se 20 jovens de boa saúde que fumaram maconha, pelo menos, 4 dias por semana, durante um mínimo de 6 semanas, sem usar qualquer outra droga durante esse período. Para comparação, usou-se um grupo de controle de 20 jovens que nunca tinham fumado maconha. A medida de sexua-

lidade usada foi o nível de hormônio do sexo masculino (testosterona) no sangue.

Fazendo os índices 1 e 2 corresponderem aos grupos de maconha e não-maconha, respectivamente, o experimento forneceu os dados seguintes:

$$\bar{x}_1 = 416 \quad \bar{x}_2 = 742 \quad s_1 = 152 \quad s_2 = 130$$

(a) 4 — Testar a hipótese $H_0 : \mu_1 = \mu_2$.

Como se presume que as amostras sejam bastante grandes para usarmos métodos de grandes amostras, basta calcular

$$\sigma_{\bar{x}_1 - \bar{x}_2} \doteq \sqrt{\frac{(152)^2}{20} + \frac{(130)^2}{20}} = 44{,}7$$

dessa forma

$$z = \frac{416 - 742}{44{,}7} = -7{,}3$$

Este valor é altamente significativo; portanto, é bastante certo que o uso habitual da maconha diminui os impulsos sexuais enquanto medido pela quantidade de testosterona no sangue.

(b) 6 — Teste a mesma hipótese usando métodos de pequenas amostras. A aplicação da fórmula (6) nos dá

$$t = \frac{416 - 742}{\sqrt{19(152)^2 + 19(130)^2}} \sqrt{\frac{20 \cdot 20 \cdot 38}{40}} = -7{,}4$$

Como o valor crítico de t baseado em 38 graus de liberdade é muito próximo de 2,02 como se vê na Tabela V, tal resultado é altamente significativo. Deve-se observar que o método das pequenas amostras difere muito pouco do método de grandes amostras para este problema. Isto implica que $n_1 = 20$ e $n_2 = 20$ sejam suficientemente grandes para justificar o uso do método das grandes amostras.

Exemplo 5. O *Time* de 22 de outubro de 1973 relatou um estudo realizado em Chapel Hill, na Carolina do Norte, para determinar se havia algo de verdade na fama que têm os orientais de se embriagarem fortemente, ao tomarem bebidas alcoólicas. O estudo consistiu em comparar 24 americanos de ascendência oriental que viviam em Chape Hill e que ingeriam bebidas alcoólicas, com 24 americanos de ascendência européia que tinham hábitos de bebida

semelhantes. Todos receberam fortes doses de bebidas e foram depois observados. Eis os resultados:

17 dos 24 orientais ficaram totalmente embriagados.

3 dos 24 ocidentais ficaram totalmente embriagados.

Teste a hipótese $H_0 : p_1 = p_2$, onde estas probabilidades correspondem aos dois grupos.

Combinando as duas amostras tem-se a estimativa $\hat{p} = 20/48 = 5/12$. Dessa forma, uma boa estimativa do desvio-padrão de $\hat{p}_1 - \hat{p}_2$ é dada por

$$\sigma_{\hat{p}_1 - \hat{p}_2} \cong \sqrt{\left(\frac{5}{12}\right)\left(\frac{7}{12}\right)\left[\frac{1}{24} + \frac{1}{24}\right]} = 0{,}142$$

Donde

$$z = \frac{\frac{17}{24} - \frac{3}{24}}{0{,}142} = 4{,}1$$

Este resultado é significativo e, portanto, parece bastante certo que os orientais têm uma fama merecida de sofrerem mais que os ocidentais ao ingerirem bebidas alcoólicas fortes.

Exemplo 6. Os valores seguintes foram obtidos realizando-se amostragens de uma população normal com $\mu = 3{,}0$ e $\sigma = 0{,}5$. Dois diferentes pesquisadores A e B colheram 25 amostras cada um. Com estes dados resolva os problemas seguintes.

(a) 2 — Combine os dados em um conjunto e teste a hipótese $H_0 : \mu = 3{,}0$ contra $H_1 : \mu \neq 3{,}0$.

(b) 2 — Construa um gráfico de controle para esses 50 valores observacionais.

(c) 4 — Teste a hipótese de que $\mu_A = \mu_B$ usando o fato de que $\sigma_A = \sigma_B = \sigma = 0{,}5$.

(d) 4 — Resolva a parte (c) sem supor que os desvios-padrão sejam conhecidos, isto é, supondo que os desvios-padrão sejam desconhecidos.

(e) 6 — Usando os dados de A e a distribuição t, teste a hipótese $H_0 : \mu_A = 3{,}3$ contra $H_1 : \mu < 3{,}3$.

(f) Usando a distribuição t teste a hipótese H_0: $\mu_A = \mu_B$ e compare seu valor de t com o valor z da parte (d).

A	3,5	3,4	2,8	2,6	2,3	2,4	2,9	2,9	3,1	3,3	3,9	3,0	2,9	3,0	2,5
B	2,9	3,6	3,1	2,5	2,6	3,6	2,9	2,1	2,6	2,5	2,3	2,6	3,3	3,9	2,7

A	2,7	3,9	2,6	2,8	3,2	3,6	2,7	3,0	3,8	3,1
B	2,4	2,8	3,0	3,3	3,7	3,3	3,1	2,7	2,4	3,3

(a) Cálculos de todos os 50 valores dão $\bar{x} = 2{,}98$; dessa forma

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{2{,}98 - 3{,}00}{\frac{0{,}5}{\sqrt{50}}} = 0{,}28$$

Como $|z| < 1{,}96$, aceita-se H_0.

(b) Uma faixa de três desvios-padrão terão os limites $\mu \pm 3\sigma = 3{,}0 \pm 1{,}5 = (1{,}5,\ 4{,}5)$. Esta faixa inclui todos os 50 pontos, desde que não haja valores menores que 1,5 e maiores que 4,5.

(c) Os cálculos dão $\bar{x}_A = 3{,}04$ e $\bar{x}_B = 2{,}93$; dessa forma

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\sigma^2\left(\frac{1}{25} + \frac{1}{25}\right)}} = \frac{0{,}11}{\sqrt{0{,}25\left(\frac{1}{25} + \frac{1}{25}\right)}} = 0{,}78$$

Como $z < 1{,}96$, aceita-se H_0.

(d) Cálculos adicionais dão $s_A^2 = 0{,}202$ e $s_B^2 = 0{,}205$; dessa forma

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{25} + \frac{s_B^2}{25}}} = \frac{0{,}11}{\sqrt{\frac{0{,}202}{25} + \frac{0{,}205}{25}}} = 0{,}86$$

Como $|z| < 1{,}96$, aceita-se H_0.

(e) Aqui $\bar{x} = 3{,}04$, $s^2 = 0{,}202$, e $\nu = 24$; dessa forma

$$t = \frac{3{,}04 - 3{,}30}{\sqrt{0{,}202}}\sqrt{25} = -2{,}89$$

Como $t < 1{,}7109$, rejeita-se H_0. Esta é a decisão correta porque $\mu_A = 3{,}0$.

(f) Aqui

$$t = \frac{3{,}04 - 2{,}93}{\sqrt{4{,}86 + 4{,}91}}\sqrt{\frac{(25)\ (25)\ (48)}{25 + 25}} = 0{,}86$$

Como $|t| < 2{,}0639$, aceita-se H_0. Os valores de t e z são os mesmos, a este grau de precisão de cálculo. Para amostras tão grandes como estas há muito pouca diferença entre os dois testes.

EXERCÍCIOS

Seção 1

1 — Num caso judicial no qual um indivíduo está sendo acusado de roubo, quais são os dois tipos de erro? Que tipo de erro é considerado mais importante pela sociedade?

2 — Dê exemplo de uma hipótese para a qual o erro do tipo II seria considerado muito mais sério que o erro do tipo I.

3 — Suponhamos que o leitor concorde em rejeitar uma hipótese se dois lances de uma moeda honesta produzirem duas caras. Quais seriam os tamanhos dos tipos de erro?

4 — Lança-se uma moeda duas vezes. Seja **x** o número de caras obtidas. Consideremos a hipótese $\mathbf{H_0} : \mathbf{p} = 0{,}5$ e a alternativa $\mathbf{H_1} : \mathbf{p} = 0{,}7$, onde **p** é a probabilidade de se obter cara e suponhamos que se vá testar esta hipótese por meio do valor de **x**. As distribuições de **x** sob $\mathbf{H_0}$ e $\mathbf{H_1}$ podem ser obtidas por meio da fórmula (1) da distribuição binomial que aparece no Capítulo 5. Suponhamos que se escolha $\mathbf{x} = 2$ como região crítica para o teste. Calcule os tamanhos dos dois tipos de erro.

5 — No problema 4 suponhamos que $\mathbf{x} = 0$ tenha sido escolhido como região crítica para o teste. Calcule agora os tamanhos dos dois tipos de erro e compare seus resultados com os do problema 4. Comente estas duas escolhas da região crítica.

Seção 2

6 — Dado $\bar{x} = 82$, $\sigma = 16$, e $n = 100$, teste a hipótese de $\mu = 86$.

7 — Dado $\bar{x} = 82$, $\sigma = 16$, e $n = 25$, teste a hipótese de $\mu = 86$.

8 — Um comprador de tijolos julga que a qualidade dos tijolos está deteriorando. De experiência passada sabe-se que a média de resistência ao esmagamento de tais tijolos é de 400 libras com um desvio-padrão de 20 libras. Uma amostra de 100 tijolos deu uma média de 395 libras. Teste a hipótese de que a qualidade média não se tenha alterado contra a alternativa de que se tenha deteriorado.

9 — Muitos anos de experiência com um exame de inglês para admissão à universidade forneceu a nota média de 64 com um desvio-padrão de 8. Todos os estudantes de certa cidade, em número de 55, obtiveram uma nota média de 68. Podemos estar certos de que os estudantes desta cidade sejam superiores em inglês?

10 — Um fabricante de linha de pescar afirma que sua linha do teste de 5 libras resiste ao teste de 8 libras. É correta sua afirmativa se uma amostra de tamanho igual a 50 deu $\bar{x} = 7$ libras e $s = 1{,}4$ libras?

11 — Construa um gráfico de controle para $\bar{x}$ para os dados seguintes sobre o tempo de retardo de fusíveis, sendo tomadas amostras de 5 cada hora. Cada conjunto de 5 foi arranjado por ordem de magnitude. Estimar μ calculando a média de todos os dados, e estimar $\sigma_{\bar{x}}$ estimando primeiro σ por meio de **s** calculado para todos os 60 valores. Dizer se parece existir controle aqui.

42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	91	78
75	68	80	69	57	75	72	27	39	113	93	94
78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

12 — Tome uma amostra de tamanho 25 de uma tabela de números aleatórios de 1 dígito (Tabela XI do apêndice) e teste a hipótese de que $\mu = 4{,}5$. Use o fato de que $\sigma = 2{,}87$ para esta distribuição e de que $\bar{x}$ pode ser tratado como uma variável normal. Traga o seu resultado para a classe. Aproximadamente 95 por cento da classe deverá aceitar essa hipótese porque ela é verdadeira.

13 — O "**Journal of the Market Research Society**", de abril de 1974, relatou que a Heinz Company desde 1971 realizou 50 testes de paladar com vários tipos de sopa. Cada teste consistia em selecionar 300 donas-de-casa, dividindo-as aleatoriamente em 2 grupos e pedindo-lhes para avaliarem duas sopas. Uma das sopas era da Heinz, a outra, do competidor. A um grupo apresentava-se primeiro a sopa Heinz e a outro essa sopa era apresentada em segundo lugar. Para determinar se a ordem de apresentação favorecia a sopa apresentada primeiro, as notas médias para a sopa apresentada em primeiro lugar e também para a sopa apresentada em segundo lugar eram calculadas para os dois grupos e registrava-se a diferença dessas médias. A escala de avaliação usada ia de 1 a 6 correspondendo a uma amplitude que ia de muito pobre a excelente. Cada experimento deu duas dessas médias, uma para o grupo que provou a sopa Heinz em primeiro lugar e a outra para o grupo que a provou em segundo lugar. Se a ordem do teste de paladar não tivesse influência nas notas do mesmo, essas médias deveriam variar em torno de zero. Os resultados dos 50 experimentos são dados no quadro seguinte.

Diferença da Contagem Média $\geq$	0,6	0,5	0,4	0,3	0,2	0,1	0	− 0,1
Número de Sopas	6	9	11	24	26	13	9	2

Como neste artigo não aparecem desvios-padrão, trate as diferenças de nota média como valores de uma variável aleatória **x** com freqüências correspondentes dadas pela segunda linha do quadro. Suponha também que a entrada $\geq$ 0,6 possa substituir-se por 0,6 quando se calcular $\bar{x}$ e **s**. Usando os dados desse modo, teste a hipótese de que a média de **x** seja zero. Qual a implicação desse resultado sobre testes de produtos desse tipo?

Seção 3

14 — Se você lançar um dado 240 vezes e obtiver 52 **seis**, concluiria que o dado favorece o número **seis**?

15 — A experiência passada mostrou que 40 por cento dos estudantes ficaram reprovados num vestibular de inglês. Se ficarem reprovados 56 de 120 estudantes de certa cidade, poderemos concluir com justeza que os estudantes dessa cidade são inferiores em inglês?

16 — Um biólogo preparou um inseticida destinado a matar 50 por cento de certo tipo de inseto. Se, ao pulverizar 200 de tais insetos, matou 120, você pode concluir que a mistura era satisfatória?

17 — Qual a diferença entre dizer que uma moeda é honesta para efeitos práticos e dizer que é honesta em sentido matemático?

18 — Explique por que você hesitaria, por considerações práticas, em rejeitar a hipótese de que uma moeda seja não viciada, se você a lançou 1 000 vezes e obteve 535 caras?

19 — Os dados seguintes se referem à percentagem defeituosa diária de peças de uma produção de aproximadamente 1 000 unidades diárias. Construa um gráfico de controle e indique se parece existir controle aqui. Os dados se referem à percentagem (não proporção) de peças defeituosas e devem ser lidos segundo as linhas. Estimar **p** pelo cálculo da média desses valores de amostra.

2,2	2,3	2,1	1,7	3,8	2,5	2,0	1,6	1,4	2,6	1,5	2,8	2,9	2,6	2,5
2,6	3,2	4,6	3,3	3,0	3,1	4,3	1,8	2,6	2,1	2,2	1,8	2,4	2,4	1,6
1,7	1,6	2,8	3,2	1,8	2,6	3,6	4,2							

Seção 4

20 — Dadas duas amostras aleatórias de tamanho 100 cada, de duas populações normais com valores amostrais $\bar{x}_1 = 20$, $\bar{x}_2 = 22$, $s_1 = 8$, $s_2 = 7$, teste a hipótese de $\mu_1 = \mu_2$.

21 — Dois conjuntos de 50 crianças de escola primária foram ensinados a ler por 2 métodos diferentes. Após o término da instrução um teste de leitura deu os seguintes resultados: $\bar{x}_1 = 73{,}4$, $\bar{x}_2 = 70{,}2$, $s_1 = 9$, $s_2 = 10$. Teste a hipótese de que $\mu_1 = \mu_2$.

22 — Num experimento industrial certa tarefa foi realizada por 40 homens conforme o método I e por 50 homens conforme o método II. Os resultados do experimento forneceram os seguintes dados sobre a duração requerida para se completar a tarefa: $\bar{x}_1 = 54$ minutos, $\bar{x}_2 = 58$ minutos, $s_1 = 6$ minutos, $s_2 = 9$ minutos. Teste a hipótese de $\mu_1 = \mu_2$.

23 — Um lote de 100 perus foi dividido em 2 grupos de 50 cada um e alimentados com duas rações diferentes. No fim do experimento obteve-se a seguinte informação: $\bar{x}_1 = 15{,}2$, $s_1 = 2{,}1$, $\bar{x}_2 = 14{,}9$, $s_2 = 2{,}0$ (unidades de peso). Teste a hipótese de que $\mu_1 = \mu_2$.

24 — No estudo dos efeitos de fortes doses de vitamina C sobre os resfriados, mencionado na seção 5, conservou-se um registro do número de resfriados contraídos por cada participante durante o período experimental. O quadro de valores que segue fornece parte desse registro, em forma levemente alterada.

	Número de Indivíduos	Número Médio de Resfriados	Desvio-padrão do Número de Resfriados
Grupo vitamínico	407	1,38	1,23
Grupo submetido a remédio só de ação psicológica	411	1,48	1,14

Use esses dados para testar a hipótese de que doses de vitamina C não tenham efeito sobre o número médio de resfriados contraídos por indivíduos da população deste experimento.

Seção 5

25 — Num inquérito feito entre estudantes universitários, 44 de 200 rapazes de um grêmio favoreceram certa proposição, acontecendo o mesmo com 51 de 300 não pertencentes ao grêmio. Há uma diferença real de opinião sobre esta proposição?

26 — Num inquérito sobre audiência de televisão numa cidade, 56 de 200 homens não gostavam de certo programa, ao passo que 75 de 300 mulheres não o apreciavam. Há aqui uma diferença real de opinião?

27 — Em uma seção de uma cidade, 62 de 450 contribuintes negligenciaram o pagamento dos impostos, enquanto em outra seção houve 40 de 500 negligentes. Teste para ver se a taxa de negligência é a mesma para as duas seções da cidade.

28 — Um teste com 200 jovens e 400 adultos mostrou que 50 dos jovens e 64 dos adultos eram motoristas descuidados. Use esses dados para determinar qual deveria ser o tamanho da diferença entre $p_1 - p_2$ se se quisesse testar a hipótese de que a diferença seria, pelo menos, deste tamanho contra a alternativa de que ela deveria ser menor.

29 — Duas classes do mesmo tamanho são ensinadas pelo mesmo professor. Para uma delas o sistema de aprendizado é o de uma classe normal; a outra está junto à primeira, mas recebe as lições do professor através de um circuito fechado de televisão. Discuta fatores possíveis que poderiam impedir uma comparação estatística válida com relação ao uso dos dois métodos de aprendizagem.

30 — Um inquérito do Instituto Gallup sobre os hábitos de bebida dos americanos, relatado no **Los Angeles Times**, de 1974, obteve os dados relacionados abaixo. As percentagens estão baseadas numa amostra de 1 543 adultos tomados de todo o país. Pelo termo "bebida alcoólica" queremos significar bebidas como licor, vinho e cerveja. As três últimas percentagens correspondem aos grupos de idade indicados.

	Nacional	Homens	Mulheres	<30	30–49	>49
Percentagem que usa bebidas alcoólicas	68	76	61	79	75	54

(a) Supondo que a amostra seja mais ou menos igualmente dividida entre homens e mulheres, teste a hipótese de $p_1 = p_2$, onde p_1 representa a proporção de pessoas do sexo masculino e p_2 a proporção de pessoas do sexo feminino que ingerem bebidas alcoólicas.

(b) Supondo que a hipótese seja mais ou menos igualmente dividida entre os três grupos de idade, teste a hipótese segundo a qual não há diferença entre o grupo de idade < 30 e o de 30 a 49.

É óbvio, sem testes, que o grupo mais velho tem uma percentagem muito menor de alcoólatras do que qualquer dos dois outros grupos.

Seção 6

31 — Dado que x é normalmente distribuído e dados valores amostrais $\bar{x} = 42$, $s = 6$, $n = 20$, teste a hipótese de $\mu = 44$, usando a distribuição t.

32 — Os seguintes dados fornecem o aumento em peso de 15 frangos alimentados com uma ração especial para aumento de peso, por um período de um mês. A experiência passada para uma dieta-padrão, em igual período, forneceu um ganho médio de 9 unidades de peso. Teste a hipótese usando um teste unicaudal, de que a nova ração não é um melhoramento. Os ganhos foram: 12,5, 9,6, 10,8, 8,1, 11,4, 10,7, 9,0, 8,4, 7,5, 11,0, 8,7, 11,0, 10,3, 8,5, 9,0.

33 — Dadas duas amostras aleatórias de tamanhos 10 e 12 de 2 populações normais independentes com $\bar{x}_1 = 20$, $\bar{x}_2 = 24$, $s_1 = 7$, $s_2 = 6$, teste por meio da distribuição **t** a hipótese de $\mu_1 = \mu_2$ supondo que $\sigma_1 = \sigma_2$.

34 — Os dados seguintes dão os aumentos de 20 ratos, dos quais metade recebeu sua proteína de amendoim cru e a outra metade a recebeu de amendoim torrado. Testar por meio da distribuição **t** se o fato de torrar os amendoins teve efeito sobre o seu valor em proteínas.

Cru	61	60	56	63	56	63	59	56	44	61
Torrado	55	54	47	59	51	61	57	54	62	58

35 — Os dados seguintes dão os ganhos em peso de 24 animais experimentais, separados em 2 grupos e alimentados com rações diferentes. Teste a hipótese de $\mu_1 = \mu_2$, supondo que $\sigma_1 = \sigma_2$.

x_1	66	64	52	48	42	44	54	58	62	74	58	54
x_2	82	50	74	58	52	60	68	50	58	72	58	66

36 — Os dados seguintes dão os efeitos de corrosão em vários rolos para tubos de aço revestidos e não revestidos. Tomando diferenças de pares de valores, testar por meio da distribuição **t** a hipótese de que a média de tais diferenças seja zero. Este é um teste de pares combinados.

Não revestidos	42	37	61	74	55	57	44	55	37	70	52	55
Revestidos	39	43	43	52	52	59	40	45	47	62	40	27

Geral

37 — Fez-se um estudo em 1965, relatado no **Journal of Pediatrics,** para determinar se era possível reconhecer sintomas prematuros de obesidade adolescente. Analisou-se uma amostra de 158 meninas obesas com média de idade de 15 anos. Tal análise dizia respeito a várias características físicas da primeira infância. Analisou-se também um grupo experimental de 94 meninas não obesas, de idade e condições sócio-econômicas semelhantes. O quadro seguinte dá as médias amostrais e os desvios-padrão para duas características dos 2 grupos.

	Grupo Obeso	Grupo não obeso
Peso ao nascer	$\bar{x}_1 = 7{,}04$, $s_1 = 1{,}2$	$\bar{x}_2 = 7{,}19$, $s_2 = 0{,}9$
Peso a um ano de idade	$\bar{x}_1 = 23{,}3$, $s_1 = 2{,}8$	$\bar{x}_2 = 21{,}9$, $s_2 = 3{,}0$

(a) Use estes dados para testar a hipótese de $H_0 : \mu_1 = \mu_2$ para peso de nascimento.

(b) Use estes dados para testar a hipótese $H_0 : \mu_1 = \mu_2$ para peso de um ano.

(c) Que conclusões podem ser retiradas dos resultados em (a) e (b)?

38 — Um grupo de doutores da universidade da Pennsylvania efetuou um experimento, referido no **Los Angeles Times** de 3 de janeiro de 1973, sobre os efeitos colaterais de vários medicamentos comuns antidistônicos. No estudo estavam em jogo 166 pacientes com sintomas de tensão e ansiedade de brandos a moderados. Foram divididos em quatro grupos por seleção aleatória e se pediu que relatassem quaisquer sintomas colaterais como vertigem, sonolência, confusão e náusea. Os resultados do experimento aparecem no quadro seguinte.

Medicação	Aspirina	Compoz	Remédios só de ação psicológica	Libritab
Número de Pacientes	38	39	35	36
Número dos que têm efeitos colaterais	3	18	4	14

Teste a hipótese que diz não haver diferença entre as proporções de efeitos colaterais para aspirina e Compoz.
Este estudo revelou que nem a aspirina nem o Compoz foi efetivo no alívio da ansiedade; contudo, a droga Libritabs, freqüentemente prescrita para aliviar ansiedade, o foi.

39 — Um grupo de médicos do departamento de Pediatria da Escola de Medicina da Universidade de Washington realizou um estudo sobre os possíveis efeitos do alcoolismo da mãe sobre a saúde do filho. Esse estudo apareceu na publicação **Lancet** de 1 de junho de 1974. Estavam em jogo 23 mulheres alcoólatras juntamente com 46 não alcoólatras escolhidas para serem cotejadas às alcoólatras com relação a características tais como idade, educação, status econômico, raça e status conjugal. Os dados que seguem representam só uma pequena parte do estudo. Embora houvesse 23 crianças de mães alcoólatras neste estudo, 4 delas morreram no espaço de uma semana e 6 outras se perderam nos estudos seguintes durante os 7 anos de observação. Só uma das crianças nascidas de mãe não alcoólatra morreu na

primeira semana após o nascimento e todas as demais permaneceram em observação durante os 7 anos do estudo. Esta parte do estudo refere-se aos quocientes de inteligência das crianças com idade de 7 anos.

	Mães Alcoólatras	Mães Não-alcoólatras
Tamanho da Amostra	13	45
Número de crianças com QI abaixo de 80	6	5

Teste a hipótese de que a percentagem de crianças com um QI menor que 80 seja a mesma para os dois grupos.

40 — Faça cada um dos membros da classe representar graficamente 10 pares de números aleatórios de 1 dígito (Tabela XI no Apêndice). O segundo número deve ser subtraído do primeiro em cada par. Cada estudante deve trazer para a classe a soma dos quadrados dessas diferenças juntamente com a soma das diferenças. Combinando os resultados da classe, uma estimativa de σ_z^2, na qual $\mathbf{z} = \mathbf{x} - \mathbf{y}$, pode ser obtida e comparada com o valor esperado da fórmula (4), a saber: $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 = (2{,}87)^2 + (2{,}87)^2 = 16{,}5$. Aqui **x** e **y** indicam respectivamente o primeiro e o segundo dígito aleatório de cada par. De trabalho anterior foi encontrado que $\mu_x = 4{,}5$ e $\sigma_x = 2{,}87$.

41 — Selecione dois conjuntos de 20 números aleatórios de 1 dígito e teste a hipótese (verdadeira aqui) pela qual as duas médias populacionais são iguais. Trabalhe primeiro com o método de grandes amostras, usando o fato de $\sigma = 2{,}87$; depois trabalhe com a distribuição **t** de Student.

42 — Os dados da Tabela XII do Apêndice podem ser usados para se construir vários problemas de testes de hipótese. Por exemplo, os dados da variável $\mathbf{x}_3$ (escore total) podem ser divididos em duas partes e pode-se fazer um teste para se observar se as duas partes possuem as mesmas médias teóricas. Como estes dados vêm de diferentes seções do mesmo curso, estas médias não precisam ser as mesmas.

9

CORRELAÇÃO E REGRESSÃO

1 — CORRELAÇÃO LINEAR

Os métodos estatísticos apresentados até aqui foram todos eles referentes à variável isolada x e sua distribuição de freqüências. Em particular os dois capítulos precedentes ligaram-se à estimativa e teste de hipóteses sobre os parâmetros das distribuições de freqüências binomial e normal da variável. Vários problemas no trabalho estatístico, todavia, envolvem muitas variáveis. Este capítulo dedica-se à explicação de uma das técnicas para lidar com dados que possuem duas ou mais variáveis. Dá-se ênfase a duas variáveis, mas o emprego dos métodos pode-se estender a mais de duas.

Em alguns problemas, as diversas variáveis são estudadas simultaneamente para se ver como elas estão inter-relacionadas; em outros há uma variável de particular interesse e as demais variáveis são estudadas pela sua possível ajuda em trazer esclarecimento sobre essa variável particular. Essas duas classes de problemas estão comumente associadas aos nomes de *correlação* e *regressão* respectivamente. A correlação será discutida primeiro.

Um problema de correlação surge quando um indivíduo se pergunta se há alguma relação entre um par de variáveis que lhe interessam. Por exemplo, há alguma relação entre fumo e doença do coração, entre sensibilidade para a música e vocação para a ciência, entre rádio-recepção e perturbações solares, entre beleza e inteligência?

Com o propósito de ilustrar como se deve proceder ao estudar a relação entre duas variáveis consideremos os dados do Quadro 1, que consiste nas médias escolares do colégio e do primeiro ano do

ciclo universitário. As médias do colégio são denotadas por x e da universidade por y.

A investigação da relação de duas variáveis, tais como estas, usualmente começa com uma tentativa para descobrir a forma aproximada dessa relação, representando-se graficamente os dados como pontos no plano x, y. Tal gráfico chama-se *diagrama de dispersão*. Por meio dele pode-se prontamente verificar se existe alguma relação pronunciada e, em caso positivo, se a relação pode ser tratada como aproximadamente linear. O diagrama de dispersão para os 30 pontos obtidos dos dados do Quadro 1 aparece na Figura 1.

QUADRO 1

x	y	x	y	x	y
3,0	2,4	2,9	1,9	3,1	2,8
2,4	2,6	2,7	2,2	3,3	3,2
3,7	3,0	3,7	3,1	2,7	1,8
3,6	3,9	2,7	2,6	3,5	2,7
3,8	3,6	3,3	2,8	2,9	2,1
2,9	3,0	2,8	2,7	2,7	1,7
3,5	3,1	3,1	2,4	2,9	1,7
3,0	2,8	2,8	3,0	3,2	2,3
2,3	2,2	3,0	3,3	3,4	2,6
3,0	2,9	2,2	1,8	2,5	2,7

Uma inspeção desse diagrama de dispersão mostra que há uma tendência para os pequenos valores de x se associarem aos pequenos valores de y e os grandes valores de x se associarem aos grandes valores de y. Além disso, grosseiramente falando, a tendência geral da dispersão é a de uma linha reta. Na determinação da natureza de uma tendência procura-se ver se há alguma inclinação pronunciada para que os pontos se espalhem de ambos os lados de alguma curva suave com poucas ondulações ou se eles parecem estar espalhados de ambos os lados de uma linha reta. Quer parecer aqui que uma linha reta serviria tão bem quanto uma curva moderadamente ondulada. Para variáveis como essas, seria desejável poder-se medir de alguma forma o grau em que as variáveis estão linearmente relacionadas. Com o propósito de idealizar tal medida, consideremos as propriedades que seriam desejáveis.

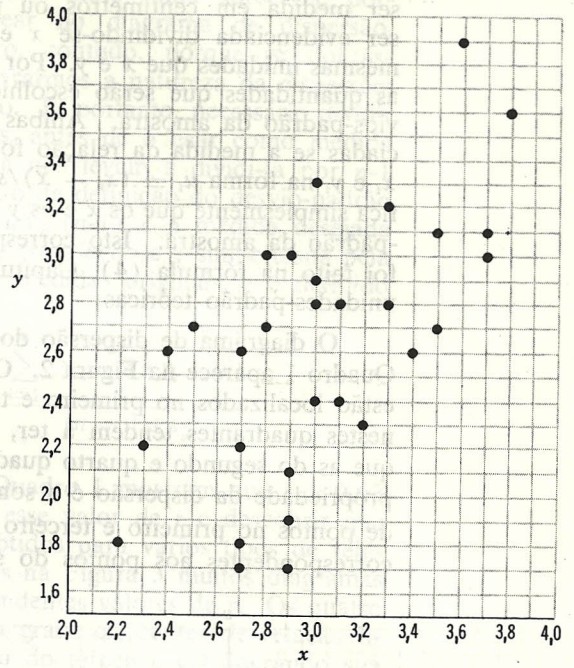


FIGURA 1. Diagrama de dispersão para médias.

Uma medida de relação deveria, certamente, independer da escolha da origem para as variáveis. O fato de o diagrama de dispersão da Figura 1 ter sido desenhado com os eixos convenientemente escolhidos a fim de passarem pelo ponto (2,0, 1,5) implica que a relação foi admitida como independente da escolha da origem. Essa propriedade pode ser utilizada usando-se os desvios das variáveis com relação à média ao invés das próprias variáveis. Isso foi feito ao definir-se o desvio-padrão, no Capítulo 2. Dessa forma usam-se as variáveis $x_i - \bar{x}$ e $y_i - \bar{y}$ em lugar das variáveis x_i e y_i na construção da desejada medida de relação. A notação x_i, y_i indica o i-ésimo par de números do Quadro 1.

Uma medida de relação deveria também independer da escala de medida usada para x e y. Assim, se as notas x e y do Quadro 1 fossem duplicadas, as relações entre as variáveis não deveriam ser afetadas. Do mesmo modo, se alguém estivesse interessado em estudar as relações entre estatura de marido e de mulher não iria querer que a medida de relação fosse depender do fato de a estatura

ser medida em centímetros ou polegadas. Esta propriedade pode ser evidenciada dividindo-se x e y por unidades que possuam as mesmas unidades que x e y. Por razões que serão apreciadas agora, as quantidades que serão escolhidas aqui são s_x e s_y, os dois desvios-padrão da amostra. Ambas as propriedades são, pois, evidenciadas se a medida da relação for construída usando-se as variáveis x_i e y_i na forma $u_i = (x_i - \bar{x})/s_x$ e $v_i = (y_i - \bar{y})/s_y$. Isto significa simplesmente que os x e os y deveriam ser medidos em unidades-padrão da amostra. Isto corresponde a fazer para amostras o que foi feito na fórmula (4), Capítulo 5, para medir uma unidade em unidades-padrão teóricas.

O diagrama de dispersão dos pontos (u_i, v_i) para os dados do Quadro 1 aparece na Figura 2. Observe-se que a maioria dos pontos estão localizados no primeiro e terceiro quadrantes e que os pontos nestes quadrantes tendem a ter, em grandeza, coordenadas maiores que as do segundo e quarto quadrantes. Uma medida simples desta propriedade da dispersão é a soma $\sum_{i=1}^{n} u_i v_i$. Os termos da soma de pontos no primeiro e terceiro quadrantes serão positivos, mas os correspondentes aos pontos do segundo e quarto quadrantes serão

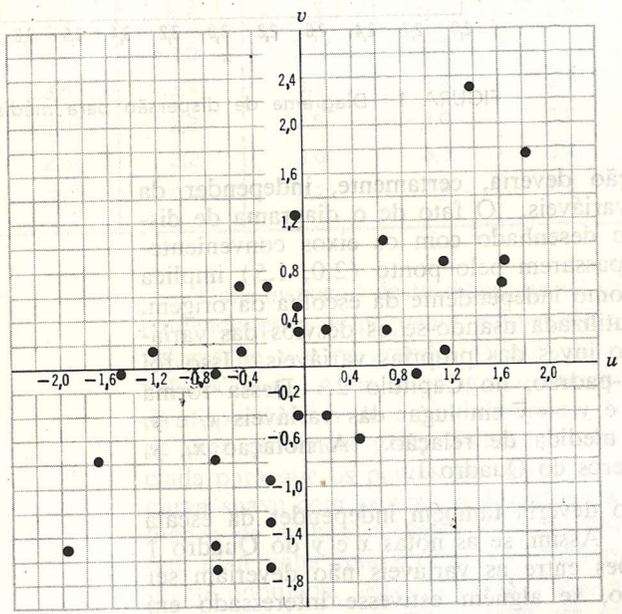


FIGURA 2. Diagrama de dispersão para médias padronizadas.

negativos. Um grande valor positivo desta soma pareceria, portanto, indicar forte tendência linear no diagrama de dispersão. Isto não é estritamente verdadeiro, contudo, porque se dobrássemos o número de pontos sem alterarmos a natureza da dispersão, o valor da soma seria quase dobrado. É, portanto, necessário dividir a soma por n, o número de pontos, antes de o usar como medida de relação. Há razões teóricas que nos levam a dividi-la por n-1 ao invés de por n aqui, como no caso da definição do desvio-padrão da amostra. A soma resultante $\Sigma u_i v_i/(n\text{-}1)$ é a medida desejada da relação. Chama-se *coeficiente de correlação* e é simbolizada pela letra r; dessa forma, em termos de medidas originais, r é definido pela seguinte fórmula:

(1) **Coeficiente de correlação**
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Os cálculos com os dados do Quadro 1 mostram que $r = 0{,}63$ para esses dados. Para interpretar esse valor de r e descobrir que valores de r serão provavelmente obtidos para vários tipos de relações entre x e y, foram apresentados na Figura 3 muitos diagramas de dispersão diferentes e os correspondentes valores de r. Os quatro primeiros diagramas correspondem a graus crescentes de relação linear. O quinto diagrama foi obtido do terceiro, considerando sua dispersão através do lado inverso da folha, sob luz forte. Isto ilustra o fato de que o valor absoluto de r mede a força da relação mas que o sinal de r é positivo se y tender a aumentar com o aumento de x e é negativo se y tender a decrescer com o aumento de x. O sexto diagrama ilustra uma dispersão na qual x e y estão intimamente ligados mas na qual a relação não é linear. Esta ilustração esclarece que r só é uma medida útil da força da relação entre duas variáveis, quando as duas variáveis estão relacionadas linearmente.

Os diagramas da Figura 3 juntamente com os valores associados de r, fazem plausíveis duas propriedades de r, isto é, que o valor de r deve satisfazer as inequações

$$-1 \leq r \leq 1$$

e que os valores de r só serão iguais a mais 1 ou menos 1 no caso de os pontos estarem totalmente em linha reta. Por métodos matemáticos pode-se demonstrar a exatidão dessas propriedades.

Multiplicando o parênteses da fórmula (1) e somando termo a termo veremos que o numerador de (1) pode ser escrito como

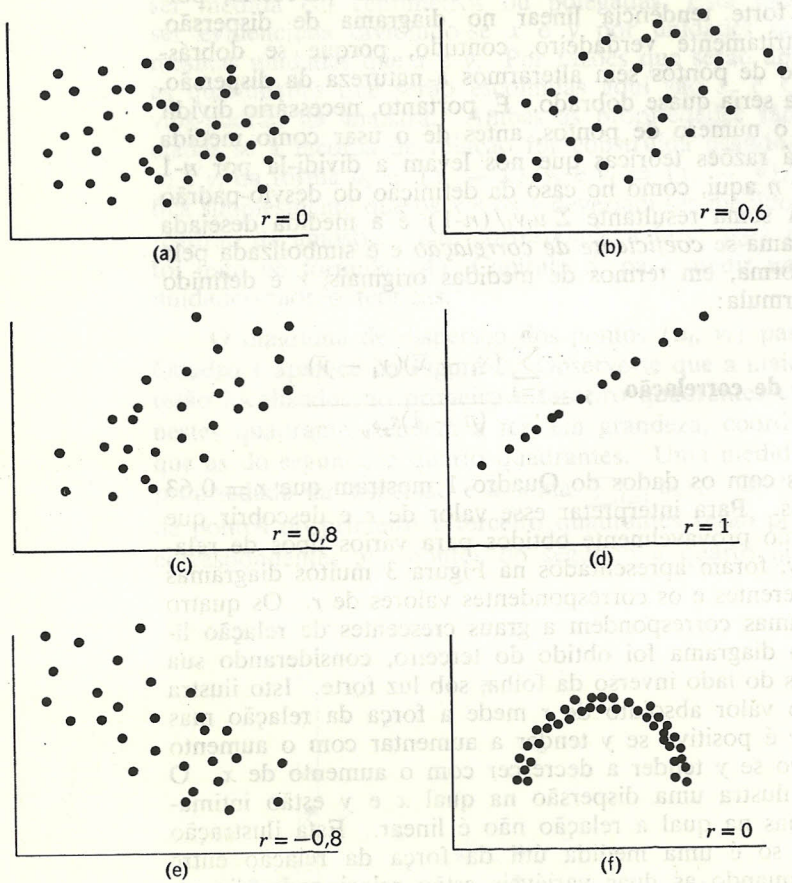


FIGURA 3. Diagramas de dispersão e os respectivos valores associados de r.

$\Sigma xy - n\bar{x}\bar{y}$, que muitas vezes é conveniente usar no cálculo do valor de r.

O valor 0,63 para a correlação entre médias de colégio e universidade é típico do tamanho do coeficiente de correlação para notas de muitos pares de assuntos referentes a colégio. A correlação de um par, como matemática e física, deveria ser um pouco mais alta que esta, ao passo que uma correlação entre matemática e teatro deveria ser mais baixa. Embora a ilustração precedente tenha produzido um valor de 0,63, alguns outros estudos mostraram um

coeficiente de correlação nas vizinhanças de 0,50 para a correlação entre médias de escolas de segundo grau e universidade. Estes valores surpreendentemente baixos são explicados em parte pelo fato de haver muita variação na qualidade do ensino ou nos critérios de avaliação das escolas. Como resultado, a média escolar não é critério seguro para a predição do sucesso escolar. Um estudo extenso feito na Inglaterra, muitos anos atrás, deu um valor de 0,51 para a correlação entre altura de pai e filho. Outro estudo produziu um valor de 0,67 para a correlação entre medidas de altura e busto de estudantes de mesma idade, de escolas elementares. Obteve-se um valor mais alto ainda, 0,80, para a correlação entre idades de marido e mulher.

Tais exemplos poderiam dar-nos uma noção do tamanho de um coeficiente de correlação provável, para determinado problema.

2 — INTERPRETAÇÃO DE r

A interpretação do coeficiente de correlação como medida da intensidade da relação linear entre duas variáveis é uma interpretação puramente matemática e está completamente isenta de qualquer implicação de causa e efeito. O fato de que as duas variáveis tendam a aumentar ou diminuir juntas não implica que uma delas tenha algum efeito direto ou indireto sobre a outra. Ambas podem ser influenciadas por outras variáveis de maneira a dar origem a uma forte relação matemática. Por exemplo, durante uma série de anos verificou-se ser de 0,98 o coeficiente de correlação entre os salários dos professores e o consumo de bebidas. Durante este período de tempo houve uma forte alta nos ordenados e salários de todos os tipos e uma tendência geral ascendente de bons tempos. Sob tais condições os salários dos professores também aumentaram. Além disso, a tendência geral ascendente nos ordenados e poder aquisitivo, juntamente com o aumento populacional, iria refletir-se num aumento da compra total de bebida. Assim, a alta correlação simplesmente reflete o efeito comum da tendência ascendente sobre duas variáveis. Os coeficientes de correlação devem ser empregados com cuidado se forem dar informações sensíveis concernentes à relação entre pares de variáveis. O sucesso com eles requer familiaridade com o campo de aplicação bem como com suas propriedades matemáticas.

Os coeficientes de correlação mostraram-se muito úteis, por exemplo, em testes psicológicos e em outros campos onde é importante determinar as relações de várias variáveis estudadas simultaneamente. Assim, correlações entre notas colegiais, notas universitárias, escores de teste de aptidão, escores de teste de vocabulário e

variáveis semelhantes permitiram a algumas instituições de ensino particular avaliar a importância relativa de tais fatores na determinação do sucesso acadêmico.

3 — FIDEDIGNIDADE DE r

O valor de r obtido para o Quadro 1 pode ser considerado como o primeiro valor de uma seqüência de valores de amostra $r_1, r_2, r_3, \ldots$, que seriam obtidos se conjuntos repetidos de dados semelhantes o fossem. Tais conjuntos de dados são considerados como tendo sido obtidos da extração de amostras aleatórias de tamanho $n = 30$ de uma população de estudantes. Se os valores de r fossem classificados numa tabela de freqüências e fosse esboçado o histograma resultante, ter-se-ia uma boa aproximação da distribuição de freqüência-limite ou teórica de r. Como no caso de outras variáveis, tais como $\bar{x}$ e $\bar{x}_1 - \bar{x}_2$, a distribuição-limite pode ser deduzida por métodos matemáticos se se fizerem os pressupostos concernentes às variáveis x e y. Se, por exemplo, supusermos que x e y seriam variáveis normais independentes, caso em que seriam necessariamente não correlacionadas, a derivação indica que a distribuição amostral desejada de r só depende de n. A Figura 4 mostra a natureza desta distribuição quando $n = 10$ e quando $n = 20$. Por meio destas distribuições pode-se determinar se um valor amostral de r é bastante grande, numericamente, para refutar a possível afirmação de que x e y são de fato variáveis não correlacionadas, isto é, se o coeficiente de correlação teórico ρ, do qual r é uma estimativa amostral, tem o valor de 0. O Quadro VI do Apêndice apresenta os valores críticos de r para estas distribuições, isto é, valores, ou seja, a probabilidade de que um valor da amostra de r exceda o valor crítico é menor que α. Os valores de α escolhidos são 0,05, 0,025 e 0,005. Se não houver relação para crer que a correlação entre x e y é positiva ou negativa, contanto que não seja zero, deve-se usar um teste bicaudal. Dessa forma, o valor crítico de r correspondente a $\alpha = 0{,}025$ fornece um teste bicaudal com um nível de significância de 0,05. Observe-se da Figura 4, por exemplo, que o valor crítico 0,025 para $n = 20$ deveria estar entre 0,4 e 0,5, que está de acordo com o valor 0,444 do Quadro VI. Para uma amostra de tamanho 20, portanto, a hipótese da não correlação entre x e y seria aceita se a grandeza de r não excedesse 0,444.

Exemplo 1. Para ilustrar como se usa a Tabela VI, consideremos sua aplicação aos dados do Quadro 1.

Como $n = 30$ e $r = 0{,}63$ para os dados do Quadro 1, segue-se do valor 0,306 da Tabela VI que se deveria rejeitar aqui uma hipótese de não correlação entre médias no secundário e na universidade. A coluna 0,05 deveria ser usada neste problema porque a alternativa à correlação zero seria certamente uma correlação positiva e portanto só a cauda direita da região crítica deveria ser selecionada.

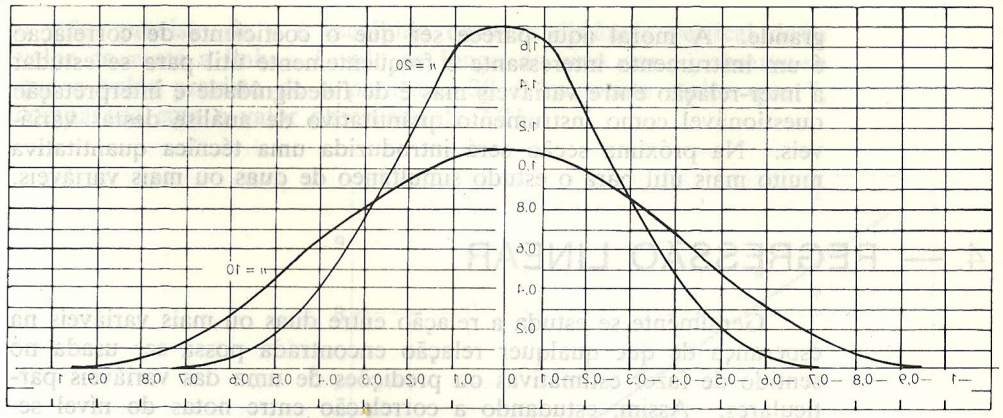


FIGURA 4. Distribuição de **r** para **p** = 0, quando **n** = 10 e **n** = 20.

Exemplo 2. Num experimento levado a cabo em 1974 na Universidade Estadual da Califórnia, em Los Angeles, para determinar o grau de relação existente entre a avaliação de um professor pelo aluno e o desempenho deste aluno no curso, várias classes com diferentes professores foram conduzidas de tal modo que todos os estudantes de um mesmo curso faziam os mesmos exames. Na conclusão do experimento, descobriu-se que o valor de r entre estas duas variáveis era 0,10. O número total de estudantes envolvidos no experimento era de 61.

Do Quadro VI depreende-se que o valor crítico 0,05 é aproximadamente 0,214; portanto, aceita-se a hipótese que afirma não haver correlação entre essas duas variáveis. Experimentos semelhantes levados a efeito por dois outros cursos com aproximadamente o mesmo número de estudantes que o experimento anterior produziram os coeficientes de correlação de $-0{,}03$ e 0,16 respectivamente. Donde, pode-se concluir com base nestes experimentos que os estudantes vão melhor num curso deste tipo com professores que não recebem alta avaliação que com professores bem avaliados pelos alunos.

Da Tabela VI parece claro que se exige uma amostra grande antes de se refutar a afirmação de que não existe correlação entre duas variáveis, a não ser, naturalmente, que o valor de r seja muito

grande. A moral aqui parece ser que o coeficiente de correlação é um instrumento interessante e freqüentemente útil para se estudar a inter-relação entre variáveis mas é de fidedignidade e interpretação questionável como instrumento quantitativo de análise destas variáveis. Na próxima seção será introduzida uma técnica quantitativa muito mais útil para o estudo simultâneo de duas ou mais variáveis.

4 — REGRESSÃO LINEAR

Geralmente se estuda a relação entre duas ou mais variáveis na esperança de que qualquer relação encontrada possa ser usada no sentido de fazer estimativas ou predições de uma das variáveis particulares. Assim, estudando a correlação entre notas do nível secundário e universitário, a intenção óbvia é usar a relação para tentar predizer o sucesso acadêmico do estudante na universidade a partir do conhecimento de suas notas no secundário. O coeficiente de correlação só mede a força do vínculo entre duas variáveis que se relacionam linearmente e é incapaz de resolver problemas de predição. Do mesmo modo, se se calculassem os coeficientes de correlação entre notas e variáveis como escores de teste de aptidão e escores de teste de vocabulário, estas correlações serviriam principalmente para indicar que variáveis se deve incluir numa função de predição para predizer sucessos acadêmicos. Os métodos destinados à manipulação de problemas de predição são conhecidos como métodos de regressão. Este capítulo discute alguns destes métodos.

Com o propósito de explicar os métodos de regressão, consideremos o problema particular de prever a colheita de feno em função da quantidade de irrigação. Os dados do Quadro 1 representam a quantidade de água aplicada e a colheita de alfafa, em toneladas por acre, numa fazenda experimental.

QUADRO 2

Água (x)	12	18	24	30	36	42	48
Colheita (y)	5,27	5,68	6,25	7,21	8,02	8,71	8,42

O gráfico desses dados é apresentado na Figura 5. Por ele, verificamos que x e y estão relacionados de modo aproximadamente linear para esta seqüência de valores de x. Uma linha reta será então ajustada a esse conjunto de pontos com o propósito de tentar prever o valor de y com base no valor de x. Tal linha foi ajustada na Figura 5. Agora, para qualquer valor de x dado, digamos $x = 30$,

o valor previsto de y é escolhido como a distância vertical deste valor de x até a linha. Lendo ao longo do eixo dos y, observe-se que o valor previsto de y para $x = 30$ é levemente maior que 7, comparado com o valor observado de 7,21.

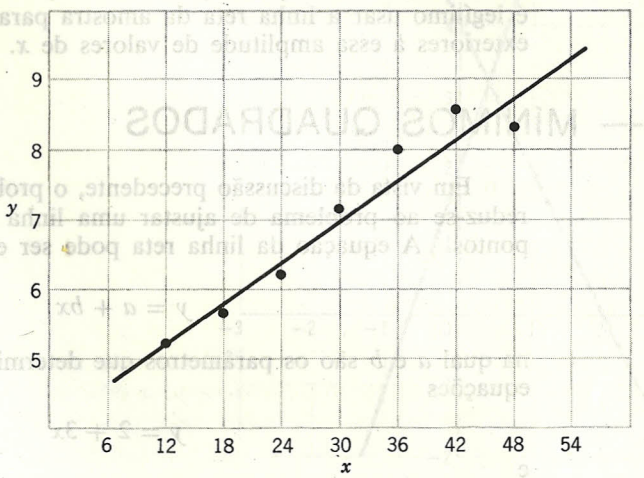


FIGURA 5. Colheita de feno como função da quantidade de irrigação.

Suponhamos, agora, que se admita que a relação entre a colheita média e a água seja estritamente linear para esta amplitude de valores de x. Significa que se este experimento tivesse sido repetido um grande número de vezes sob as mesmas condições e se houvesse sido tirada separadamente a média dos valores de y correspondentes a cada um dos sete valores de x, então essas médias forneceriam um conjunto de pontos quase exatamente em linha reta. Quanto maior o número de tais repetições, tanto melhor a precisão esperada. Essa suposição diz essencialmente que há uma linha reta teórica que expressa a relação linear entre o valor teórico médio de y e o valor correspondente de x.

Se for aceita a suposição de linearidade, pode-se esperar que o valor da linha reta da amostra de um pouco mais de 7 esteja mais próximo do valor teórico linear para $x = 30$ do que o valor observado de 7,21, porque é de supor que a linha reta da amostra, que é baseada em todos os 7 pontos experimentais, seja mais estável que um ponto isolado observado. Em vista deste raciocínio, prever-se-ia o valor teórico linear correspondente a $x = 30$ como o valor correspondente de y na linha de regressão da amostra. Pre-

visões semelhantes seriam feitas para os outros 6 valores de x. Além disso, se alguém estivesse interessado num valor intermediário de x, usaria também o valor da linha reta da amostra como o valor previsto de y para este valor de x. Visto que se está supondo que a relação seja linear apenas para esta amplitude de valores de x, não é legítimo usar a linha reta da amostra para prever os valores de y exteriores a essa amplitude de valores de x.

5 — MÍNIMOS QUADRADOS

Em vista da discussão precedente, o problema da previsão linear reduz-se ao problema de ajustar uma linha reta a um conjunto de pontos. A equação da linha reta pode ser escrita sob a forma

$$y = a + bx$$

na qual a e b são os parâmetros que determinam a linha. Assim as equações

$$y = 2 + 3x$$

e

$$y = 4 - 2x$$

determinam as duas linhas retas desenhadas na Figura 6. O parâmetro a determina o ponto onde a linha corta o eixo dos y. Assim as duas linhas da Figura 6 cortam o eixo dos y em 2 e 4 respectivamente. O parâmetro b determina a inclinação da linha. A inclinação 3 para a primeira linha significa que a linha se eleva de 3 unidades verticalmente para cada variação de unidade positiva horizontalmente. Um valor negativo como -2 significa que a linha cai 2 unidades para cada mudança positiva horizontal de 1 unidade. Visto que são necessários somente dois pontos para traçar uma linha reta, essas linhas serão prontamente desenhadas atribuindo-se dois valores diferentes a x e calculando-se os valores correspondentes de y para ter as coordenadas de 2 pontos sobre a linha. Para traçar as duas linhas da Figura 6, foram calculados os seguintes pares de pontos:

$y = 2 + 3x$

x	0	1
y	2	5

$y = 4 - 2x$

x	0	2
y	4	0

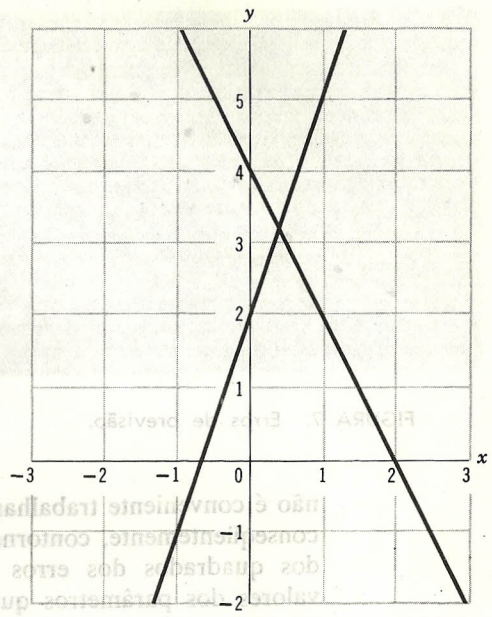


FIGURA 6. Gráficos de duas linhas retas.

Visto que o problema é determinar os valores dos parâmetros a e b de modo que a linha se vá ajustar bem a um conjunto de pontos, o problema é essencialmente o de estimar os parâmetros a e b de algum modo eficiente. Embora haja inúmeros métodos para efetuar a estimativa de tais parâmetros, o mais conhecido para os problemas da regressão é o *método dos mínimos quadrados*.

Como a linha desejada vai ser usada para fins de previsão, é razoável exigir que ela seja tal que torne pequenos os erros dessa previsão. Um erro de previsão significa a diferença entre um valor observado de y e o valor correspondente de y da linha reta. Por exemplo, o erro de precisão da Figura 5 para $x = 30$ é aproximadamente igual a $7{,}21 - 7 = 0{,}21$. Os vários erros de predição para todos os sete valores de x são mostrados na Figura 7 por meio de segmentos verticais que ligam os pontos à linha. Os pontos acima da linha dão erros positivos, os situados abaixo da linha dão erros negativos; portanto não seria bastante exigir que a soma dos erros fosse tão pequena quanto possível. Esta dificuldade pode ser evitada exigindo-se que a soma dos valores absolutos, isto é, a grandeza dos erros, seja tão pequena quanto possível. Contudo, matematicamente

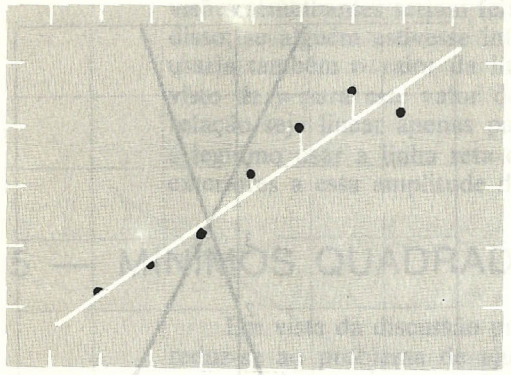


FIGURA 7. Erros de previsão.

não é conveniente trabalharmos com as somas dos valores absolutos, conseqüentemente, contorna-se a dificuldade exigindo-se que a soma dos quadrados dos erros seja tão pequena quanto possível. Os valores dos parâmetros que minimizam a soma dos quadrados dos erros determinam o que é conhecido como a *linha reta mais bem ajustada no sentido dos mínimos quadrados*. É claro do exame das Figuras 5 e 6 que se variarmos a e b de maneira conveniente poderíamos achar a equação de uma linha que se ajuste bem aos pontos da Figura 5. Mas, o problema é encontrar a linha de melhor ajustamento, de maneira sistemática e racional, e aqui é onde entra o princípio dos mínimos quadrados.

O problema de determinar os valores dos mínimos quadrados de a e b requer mais base matemática do que a que aparece neste livro; portanto, não será considerado aqui. Contudo, os resultados da aplicação dos respectivos métodos matemáticos aparecem aqui. São um tanto mais simples se a variável x for medida a partir da média amostral. Assim a equação da reta é escrita sob a forma

$$(2) \qquad y' = a + b(x - \bar{x})$$

Usa-se y' somente para distinguir o valor de y da linha reta do valor observado de y para qualquer x.

Quando a equação de uma linha reta é escrita sob a forma (2), b é ainda a inclinação da linha, mas a é agora o valor de y quando $x = x$ ao invés de ser o valor de y quando $x = 0$. As estimativas

de a e b dos mínimos quadrados para a equação da linha expressa na forma (2) são as seguintes:

$$a = \bar{y} \qquad \text{e} \qquad b = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$$

Nestas fórmulas admite-se que há n pares de valores de x e y, tais como os do Quadro 2 e que x_i e y_i indicam o i-ésimo par de valores. Embora a amplitude da soma não seja indicada no símbolo Σ, compreende-se que vai de $i = 1$ a $i = n$. Assim, para o Quadro 2, a soma seria de $i = 1$ a $i = 7$. Estes valores quando substituídos na fórmula (2) fornecem a desejada linha dos mínimos quadrados. Esta linha é comumente chamada *linha de regressão de y em x;* dessa forma os resultados precedentes podem ser escritos assim:

(3) **Linha de Regressão.** $y' = \bar{y} + b(x - \bar{x})$

onde $b = \dfrac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$

Um pioneiro no campo da estatística aplicada deu este nome à linha dos mínimos quadrados devido a alguns estudos que estava então fazendo ao estimar a extensão com que a estatura de filhos de pais altos reverte ou regressa em direção à estatura média da população. Por outras palavras, ele descobriu que os filhos de pais altos tendem a ser mais baixos que os pais, e filhos de pais baixos tendem a ser mais altos que estes últimos. Para fins de cálculo, é conveniente alterar um pouquinho a forma de b. Se multiplicarmos à parte os fatores no numerador e simplificarmos o denominador, facilmente demonstramos que b se reduz a

$$(4) \qquad b = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

O Quadro 3 ilustra o processo de cálculo para os dados do Quadro 2. Aqui a fórmula (3) foi usada para calcular b ao invés da fórmula de cálculo (4) sugerida, porque o x é aqui inteiro, a saber: 30. A fórmula (3) dá então os valores.

$$a = \frac{49{,}56}{7} = 7{,}08 \quad \text{e} \quad b = \frac{103{,}68}{1008} = 0{,}103$$

Conseqüentemente, a equação da linha de regressão é

$$y' = 7{,}08 + 0{,}103\,(x - 30)$$

QUADRO 3

x	y	$x - \bar{x}$	$(x - \bar{x})y$	$(x - \bar{x})^2$
12	5,27	−18	−94,86	324
18	5,68	−12	−68,16	144
24	6,25	− 6	−37,50	36
30	7,21	0	0	0
36	8,02	6	48,12	36
42	8,71	12	104,52	144
48	8,42	18	151,56	324
210	49,56		103,68	1008

A multiplicação do parêntese seguida da soma dos termos semelhantes dará

$$y' = 3{,}99 + 0{,}103x \tag{5}$$

Esta é a equação da linha de regressão representada na Figura 5. Algumas calculadoras eletrônicas de mão acham diretamente a equação de regressão sem a necessidade de se calcular somas de quadrados ou produtos. Este problema ilustra muito bem a diferença básica entre correlação e regressão para duas variáveis. No problema de correlação correspondente ao diagrama de dispersão da Figura 1 os dados consistiam numa amostra aleatória de tamanho 30 de uma população de estudantes. Isto significa que tanto x como y são variáveis estatísticas cujos valores são determinados só após a obtenção da amostra. No problema de regressão anterior, contudo, os valores de x foram calculados primeiro, de modo que só os valores de y foram determinados pela amostra. Ora, a técnica dos mínimos quadrados de ajustamento de uma linha a um conjunto de pontos pode ser aplicada quer fixemos de antemão os valores de x quer o obtenhamos de amostras aleatórias; dessa forma os métodos de regressão poderiam ter sido aplicados a dados do tipo considerado no estudo de correlação. Por outro lado, a interpretação de r como uma medida da força da relação linear entre duas variáveis, naturalmente, não tem sentido se os valores de x forem escolhidos como desejados, porque o valor de r dependerá muito da escolha dos valores de x.

Além de serem mais flexíveis, os métodos de regressão também possuem a vantagem de serem os métodos naturais a se usar em muitas situações experimentais. O experimentador freqüentemente deseja mudar x por quantidades uniformes ao longo da amplitude de interesse para essa variável, de preferência a tomar uma amostra aleatória de valores de x. Assim, se ele desejasse estudar o efeito de um aminoácido sobre o crescimento, iria querer aumentar a quantidade desse aminoácido de uma quantia fixa, ou fator, cada vez que realizasse o experimento.

Embora um coeficiente de correlação seja útil para descrever a intensidade da relação linear entre duas variáveis, ele não é muito útil de outra forma. Se o coeficiente de correlação entre vocação para matemática e vocação para música é de 0,4 enquanto o de vocação para matemática e vocação para arte é de 0,2, pode-se concluir que a primeira associação é mais forte que a segunda, e que ambas são bastante fracas, porém nenhum esclarecimento melhor pode ser dado. Os coeficientes de correlação não se prestam prontamente a afirmações quantitativas a menos que estejam associados à regressão. Assim, a correlação é habitualmente apenas uma primeira fase do estudo da relação de duas variáveis, enquanto a regressão é a técnica básica em tal estudo.

6 — A FALÁCIA DA REGRESSÃO

O nome de regressão, comumente dado à linha dos mínimos quadrados, está associado a um erro freqüentemente cometido na interpretação de observações tomadas em dois períodos de tempo diferentes. Este erro, chamado falácia da regressão, fica mais bem explicado por meio de ilustrações.

Suponhamos que um professor estude as notas de seus alunos, num dado curso, nos seus testes das primeiras duas horas. Se ele selecionar, digamos, as cinco melhores provas do primeiro teste e calcular a nota média para aqueles cinco estudantes em ambos os testes, possivelmente descobrirá que a média do segundo teste é inferior à do primeiro. Do mesmo modo, se calcular a nota média em ambos os testes para os cinco estudantes que tiverem as notas mais baixas no primeiro teste, sem dúvida descobrirá que a média se elevou. Assim, poderia concluir que os bons estudantes estão relaxando, enquanto os estudantes fracos estão melhorando. A explicação se baseia, em parte, na reação de estudantes quanto à nota de testes e também parcialmente na variabilidade natural das notas de testes estudantis. Mesmo que os estudantes não variassem de um teste para outro em seus hábitos de estudo e em seu conhecimento relativo total da matéria, a imprecisão de um teste para medir

este conhecimento iria causas variação considerável no desempenho desses estudantes, de teste para teste. Como resultado, alguns dos 5 melhores estudantes do primeiro teste poderão assim estar classificados devido a circunstâncias fortuitas; contudo, pode-se esperar que o segundo teste os traga de volta ao seu nível natural, fazendo decair com eles a média desse grupo de cinco estudantes. O mesmo tipo de raciocínio aplica-se aos cinco estudantes mais fracos no primeiro teste para dar a razão da melhora de sua média.

Há, sem dúvida, um fator psicológico operando aqui também para acentuar a *regressão* de notas altas e baixas em direção à média do grupo inteiro. É de se presumir que os estudantes fracos no primeiro teste se ponham a estudar consideravelmente mais que antes, e então aumentem as suas notas no segundo teste. Visto que os sucessos iniciais freqüentemente conduzem a um excesso de confiança, poderíamos esperar que os estudantes que tiraram notas mais altas no primeiro teste tenham facilitado um pouco, baixando então sua nota média no segundo teste. Contudo, a variação natural das notas de teste sozinha será suficiente para produzir uma quantidade razoável de regressão em direção à média do grupo.

A falácia da regressão freqüentemente ocorre na interpretação de dados de negócios. Por exemplo, comparando-se os lucros obtidos por um grupo de firmas de negócios similares, por dois anos consecutivos, poderia haver a tentação de se afirmar que as firmas de altos lucros estão-se tornando menos eficientes, enquanto as de baixos lucros aumentam a eficiência, porque o lucro médio para cada um daqueles dois grupos tenderia a deslocar-se em direção à média do grupo inteiro. As firmas com altos lucros no primeiro ano ocuparam essa posição na lista ou porque são normalmente muito eficientes ou porque têm obtido sucesso esse ano, sendo normal de eficiência inferior. Poder-se-ia esperar que o último grupo fosse apresentar lucros inferiores no segundo ano, decrescendo dessa forma a média do grupo de altos lucros no primeiro ano.

O mesmo tipo de raciocínio explicaria o aumento aparente da eficiência do grupo de baixos lucros no primeiro ano.

O estudo original da relação entre a estatura de pais e filhos que deu origem ao nome de regressão é outra ilustração desse tipo de possível erro de interpretação. Descobriu-se que o grupo mais alto do total dos homens estudados tinha filhos cuja altura média era inferior à dos pais. Descobriu-se também que o grupo mais baixo tinha filhos de altura média superior à dos pais. Como nas outras ilustrações, a explicação repousa na variação natural dos subgrupos de uma população. Como muitos homens altos vêm de famílias cujos pais são de estatura média, é provável que tais homens tenham filhos menores que eles; conseqüentemente, quando

se seleciona um grupo de homens altos não se vai esperar que os filhos de tais homens sejam tão altos quanto os pais. Há certamente fatores tais como a tendência de homens altos casarem com mulheres de estatura mais alta que a média, e o aumento constante na estatura, de geração a geração, para enfraquecer um pouco a tendência de regressão acima.

7 — ERRO-PADRÃO DA ESTIMATIVA

Após uma linha de regressão ter sido ajustada a um conjunto de pontos, geralmente é possível inspecionar o seu gráfico, tal como o da Figura 5 e observar com que precisão ele prediz os valores de y. Um processo aritmético para fazer isso é calcular os tamanhos de todos erros, $y_i - y'_i$, pela leitura no gráfico, se houver gráfico, ou por meio da fórmula (3). Uma medida útil da predição da previsão é obtida calculando-se a média dos erros de predição elevados ao quadrado. Esta média é dada pela expressão

$$\frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{n}$$

Para problemas relacionados à descoberta dos limites de confiança para os parâmetros a e b, uma divisão por $n-1$ e não por n parece melhor dividir a soma dos quadrados dos erros por $n-2$ ao invés de $n-1$, do mesmo modo que se julgou, na definição da variância da amostra ordinária, que define a amostra ordinária de variância. Valem aqui os mesmos argumentos que valeram lá. Pode-se demonstrar que o resultado é uma estimativa não tendenciosa de σ^2,

QUADRO 4

y	y′	y − y′	(y − y′)²
5,27	5,23	0,04	0,0016
5,68	5,84	−0,16	0,0256
6,25	6,46	−0,21	0,0441
7,21	7,08	0,13	0,0169
8,02	7,70	0,32	0,1024
8,71	8,32	0,39	0,1521
8,42	8,93	−0,51	0,2601
			0,6028

$$S_e = \sqrt{\frac{0,6028}{5}} = 0,35$$

onde σ^2 é a variância dos erros de predição quando os valores de y'_i são os de uma linha de regressão teórica e não os de uma linha de regressão amostral. A raiz quadrada da expressão resultante é simbolizada por s_e e chama-se *erro-padrão de estimativa.* Assim, em termos das variáveis originais,

(6) **Erro-padrão da estimativa.**

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y'_i)^2}{n-2}}$$

O Quadro 4 mostra os cálculos de s_e para os dados do Quadro 2 por meio das fórmulas (5) e (6).

A não ser que os valores de y'_i sejam necessários para outros propósitos, geralmente é mais simples calcular s_e por meio da seguinte fórmula, que podemos demonstrar como equivalente à fórmula (6):

(7)
$$s_e = \sqrt{\frac{\Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i}{n-2}}$$

Deveríamos chamar a atenção para o fato de que todos os valores necessários nesta fórmula, exceto um, já estão evidenciados pelos cálculos exigidos para obter a equação da linha de regressão. A letra a aqui é a mesma da equação $y' = a + bx$; dessa forma, é preciso buscar os termos da fórmula (3) para achar a.

Se se supõe, como antes, que haja uma linha de regressão teórica, da qual a reta dos mínimos quadrados é uma estimativa e, além disso, se se supõe que os valores $y_i - y'_i$ (onde y'_i é agora o valor da linha teórica) têm uma distribuição independente e teórica com média zero e o mesmo desvio-padrão σ, então s_e dado pela (6) é uma estimativa de σ. Além disso, a suposição da distribuição normal nos permite fazer afirmações probabilísticas aproximadas sobre os erros de predição. Por exemplo, pode-se afirmar que 95 por cento dos erros de predição serão menores que $1{,}96s_e$ em grandeza. A aproximação surge porque $1{,}96\sigma$ foi substituído por sua estimativa amostral $1{,}96s_e$ e porque só se dispõe da linha de regressão da amostra; portanto, esta é uma técnica para amostras grandes. Como a maioria dos problemas de regressão envolve pequenas amostras, especialmente o tipo experimental como o do Quadro 2, esta técnica é de uso um tanto limitado. Na próxima seção discutiremos um método para amostras pequenas, no trato com problemas de regressão.

Mesmo a amostra dada pelo Quadro 2 sendo pequena demais para justificar os métodos das grandes amostras que acabamos de discutir, os resultados dos cálculos que forneceram a equação (5) e o valor 0,35 para a fórmula (6) serão usados para esboçar o tipo de gráfico que se poderia legitimamente construir se se tivesse uma amostra maior. Para este fim, basta calcular o número $1{,}96s_e = 1{,}96(0{,}35) = 0{,}7$ e depois traçar duas linhas paralelas à linha (5) que é mostrada na Figura 5, estando uma 0,7 unidades acima da linha e a outra estando 0,7 unidades abaixo dela. Estas duas linhas são mostradas na Figura 8. Da Figura 8 ver-se-á que todos

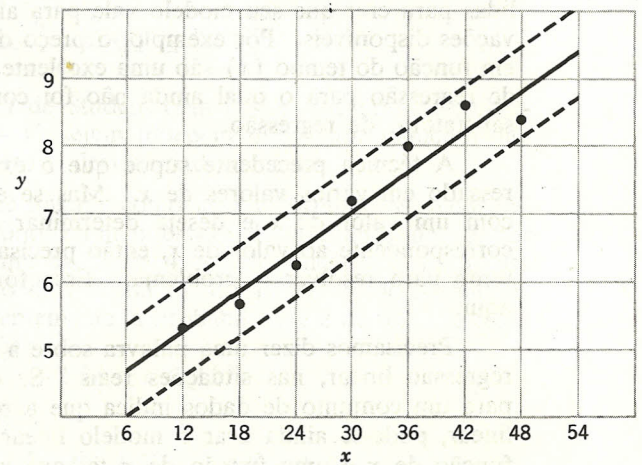


FIGURA 8. Faixa de previsão com aproximação de 95 por cento.

os sete pontos caem dentro da faixa dos 95 por cento, o que naturalmente seria de esperar. Num processo longo de experimentos semelhantes esperar-se-ia que aproximadamente 95 por cento de tais pontos caíssem dentro da faixa construída deste modo, desde que se tenha satisfeito a suposição da distribuição normal. Esta maneira geométrica de considerar o problema de predição é útil para dar ao experimentador uma idéia grosseira dos prováveis valores de y que ele obterá se experimentar os outros valores de x.

O problema de predizer um valor de y para um valor de x, além do campo dos valores observados; é consideravelmente mais difícil que a predição de um valor de x dentro do intervalo de observações. A predição de valores além do campo de observações chama-se *extrapolação,* ao passo que a predição dentro do campo

das observações chama-se *interpolação*. A dificuldade com a extrapolação é que os pressupostos necessários para justificá-la raramente se verificam na vida real. Assim, é altamente improvável que a relação entre a quantidade de irrigação e a colheita do feno continue a ser linear além do âmbito dos valores de x dados na Figura 1. Certamente não o será se o x escolhido for suficientemente grande. Além disso, na maior parte dos problemas reais de regressão não é razoável esperar que os erros de predição estejam normalmente distribuídos em torno da linha de regressão com o desvio-padrão desses erros, permanecendo constante quando x se estende além do campo de observações. A extrapolação só é uma técnica legítima quando o experimentador tem razões válidas para crer que seu modelo vale para além do campo das observações disponíveis. Por exemplo, o preço das ações de mercado (y) em função do tempo (x) são uma excelente ilustração de um modelo de regressão para o qual ainda não foi construído nenhum modelo satisfatório de regressão.

A técnica precedente supõe que o experimentador esteja interessado em vários valores de x. Mas se ele estiver preocupado só com um valor de x e deseja determinar a exatidão do y predito correspondente ao valor de x, então precisa-se de uma fórmula diferente para resolver o problema. Essa fórmula não será estudada aqui.

Precisamos dizer uma palavra sobre a utilidade dos modelos de regressão linear, nas situações reais. Se o diagrama de dispersão para um conjunto de dados indica que a relação entre x e y não é linear, pode-se ainda usar o modelo linear se pudermos achar uma função de x e uma função de y tal que a relação entre esses dois *valores funcionais* seja linear. Há, na ciência, muitas relações conhecidas que não são lineares mas que se podem tornar lineares se considerarmos as funções adequadas de x e y. Esta técnica de considerar a relação *entre as funções* de x e y, e não *entre* x e y, amplia consideravelmente o campo de aplicação dos métodos lineares.

8 — MÉTODO DAS PEQUENAS AMOSTRAS

Conforme indicado no parágrafo precedente, a fim de tornar possíveis as afirmações de probabilidade sobre os erros de previsão, é necessário admitir que existe uma linha de regressão teórica, da qual a linha dos mínimos quadrados é uma estimativa de amostra, e então fazer algumas suposições de distribuição. A linha postulada é escrita sob a forma

$$y' = \alpha + \beta(x - \bar{x})$$

O α e o β desempenham o mesmo papel para esta linha teórica que o a e o b para a linha determinada pelos pontos da amostra e dada pela fórmula (2). Assim, os valores de a e b dados pela fórmula (3) são as estimativas de α e β obtidas pelos mínimos quadrados. Não há novos problemas de estimativa ou testes de hipótese relacionados com o parâmetro α, porque, da fórmula (3), sua estimativa é y e conseqüentemente o método de pequena amostra para lidar com a média normal, explicado nos capítulos anteriores, pode ser aplicado aqui. O parâmetro de inclinação β, contudo, requer novas fórmulas. Felizmente pode ser demonstrado que a distribuição t de Student pode ser aplicada aos problemas relacionados com β. Em particular, pode ser demonstrado que a quantidade

$$(8) \qquad t = \frac{b - \beta}{s_e}\sqrt{\Sigma (x_i - \bar{x})^2}$$

possui uma distribuição t de Student com $\nu = n - 2$ graus de liberdade, desde que os $y_i - y'_i$ sejam independente e normalmente distribuídos com médias zero e desvios-padrão iguais.

Com o propósito de ilustrar como usar a fórmula (8) consideremos o problema: encontrar um intervalo de confiança de 95 por cento para β no problema da seção 4. Visto que $\nu = 5$ para esse problema, o valor crítico de t de 0,025, da Tabela V do Apêndice, será $t = 2{,}57$. Conseqüentemente, a probabilidade é de 0,95 de que t vá satisfazer as inequações

$$-2{,}57 < t < 2{,}57$$

Da (8) isso é equivalente a

$$-2{,}57 < \frac{b - \beta}{s_e}\sqrt{\Sigma (x_i - \bar{x})^2} < 2{,}57$$

Se estas inequações forem resolvidas para β, verificar-se-á que elas são equivalentes a

$$(9) \qquad b - 2{,}57\frac{s_e}{\sqrt{\Sigma (x_i - \bar{x})^2}} < \beta < b + 2{,}57\frac{s_e}{\sqrt{\Sigma (x_i - \bar{x})^2}}$$

Se 2,57 é substituído pelo valor de t conveniente, estas desigualdades podem servir como uma fórmula para encontrar os limites de confiança de β para qualquer tamanho de amostra e qualquer percentagem de limites desejada. Os cálculos anteriores deram os valores de $b = 0{,}103$ e $s_e = 0{,}35$. A última coluna do Quadro 3 fornece o valor $\Sigma (x_i - x)^2 = 1008$; dessa forma $\sqrt{\Sigma (x_i - x)^2} = 31{,}7$. Se estes valores forem substituídos na fórmula (9) eles irão se reduzir a

$$0{,}103 - 2{,}57\,\frac{0{,}35}{31{,}7} < \beta < 0{,}103 + 2{,}57\,\frac{0{,}35}{31{,}7}$$

ou

$$0{,}075 < \beta < 0{,}131$$

Como resultado, o intervalo de confiança de 95 por cento desejado para β é o intervalo de 0,075 a 0,131. Os testes de hipóteses sobre β podem ser feitos por meio da fórmula (8) da maneira usual.

7 — ILUSTRAÇÕES ADICIONAIS

Exemplo 1. Os seguintes pares de números são os escores obtidos por dez estudantes num teste de língua (x) e num teste de ciências (y).

(a) 1 — Desenhe um diagrama de dispersão e observe se a relação pode ser tratada como sendo linear. (b) 1 — Calcule o valor de r. (c) 3 — Teste a hipótese de que $\rho = 0$. (d) 5 — Ache a equação da reta dos mínimos quadrados. (e) 7 — Calcule os erros de predição. (f) 7 — Calcule o erro-padrão da estimativa. (g) 7 — Determine que percentagens de erros de predição excede $2\,s_e$.

x	56	60	64	82	76	72	74	66	64	86
y	60	68	60	74	80	84	80	72	62	82

(a) Vide diagrama da página seguinte.

(b) Os cálculos dão $\Sigma x = 700$, $\Sigma y = 722$, $\Sigma xy = 51.168$, $\Sigma x^2 = 49.840$, $\Sigma y_2 = 52.908$; dessa forma

$$\Sigma xy - n\bar{x}\bar{y} = 628,$$

$$(n-1)s_x s_y = \sqrt{(n-1)s_x^2 \cdot (n-1)s_y^2} = \sqrt{[\Sigma x^2 - n\bar{x}^2][\Sigma y^2 - n\bar{y}^2]}$$

$$= \sqrt{[840][779{,}6]} = 809{,}2; \text{ donde } r = \frac{628}{809{,}2} = 0{,}78.$$

(c) Porque se espera uma correlação positiva se não houver correlação, devemos usar um teste unicaudal. Da Tabela VI o valor crítico para $n = 10$ é 0,549. Como $0{,}78 > 0{,}549$, rejeita-se $H_0 : \rho = 0$.

(a)

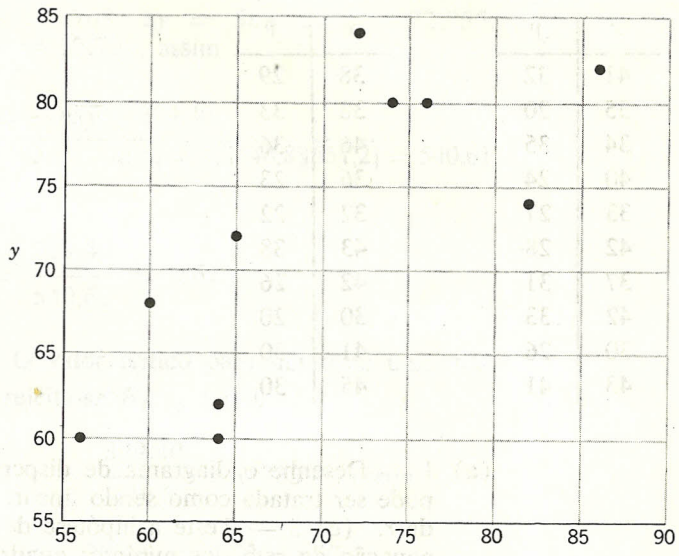


Diagrama de dispersão

(d) Usando a fórmula (4), $b = \dfrac{628}{840} = 0{,}75$; dessa forma

$$y' = 72{,}2 + 0{,}75(x - 70) = 19{,}7 + 0{,}75x$$

(e)

x	56	60	64	82	76	72	74	66	64	86
y	60	68	60	74	80	84	80	72	62	82
y'	61,7	64,7	67,7	81,2	76,7	73,7	75,2	69,2	67,7	84,2
e	−1,7	3,3	−7,7	−7,2	3,3	10,3	4,8	2,8	−5,7	−2,2

(f) $\Sigma(y-y')^2 = 310{,}1$; dessa forma $s_e = \sqrt{\dfrac{310{,}1}{8}} = 6{,}2$

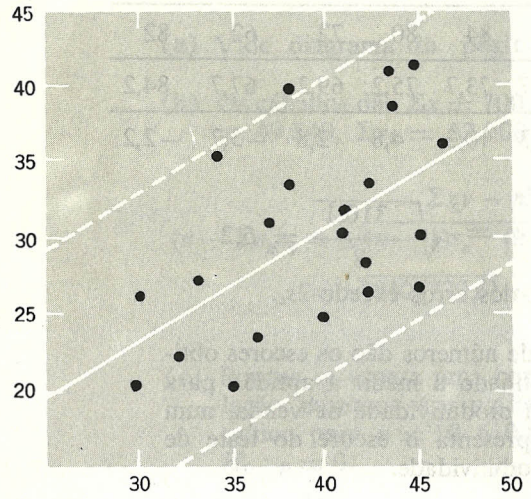
(g) $2s_e = 12{,}4$; dessa forma nenhum dos erros excede $2s_e$.

Exemplo 2. Os seguintes pares de números dão os escores obtidos por 20 vendedores num teste destinado a medir a aptidão para vendas e escores destinados a medir a produtividade de venda, num certo período de tempo. Aqui x representa o escore do teste de aplicação e y o escore do teste de produtividade.

x	y	x	y
41	32	38	29
35	20	38	33
34	35	46	36
40	24	36	23
33	27	32	22
42	28	43	38
37	31	42	26
42	33	30	20
30	26	41	30
43	41	45	30

(a) 1 — Desenhe o diagrama de dispersão e observe se a relação pode ser tratada como sendo linear. (b) 1 — Calcule o valor de r. (c) 3 — Teste a hipótese de $\rho = 0$. (d) 5 — Ache a equação da reta dos mínimos quadrados. (e) 7 — Calcule o erro-padrão de estimativa por meio da fórmula (7). (f) 7 — Trace a reta de regressão sobre o diagrama de dispersão e construa uma faixa $2s_e$ em torno dela. (g) 7 — Determine a percentagem dos pontos do diagrama de dispersão que cai dentro desta faixa. (h) 8 — Teste a hipótese de $\beta = 0$. (i) 8 — Ache um intervalo de confiança de 95 por cento para β. Seguem-se as soluções.

(a) e (f)



(b) Os cálculos dão $\Sigma x = 768$, $\Sigma y = 584$, $\Sigma xy = 22.755$, $\Sigma x^2 = 29,940$, $\Sigma y^2 = 17,704$; assim

$$\Sigma xy - n\bar{x}\bar{y} = 329,40$$

$$(n-1)s_x s_y = \sqrt{[\Sigma x^2 - n\bar{x}^2][\Sigma y^2 - n\bar{y}^2]} = \sqrt{[448,8][651,2]} = 540,61$$

donde

$$r = \frac{329,40}{540,61} = 0,61$$

(c) $n = 20$, $r = 0,61$. O valor crítico para um teste unicaudal é de 0,378; portanto rejeita-se H_0: $\rho = 0$.

(d) Usando a fórmula (4), $b = \dfrac{329,40}{448,8} = 0,73396$; assim

$$y' = 29,2 + 0,73396\ (x - 38,4) = 1,02 + 0,734x$$

Arredondando 0,73396 para 0,734 teríamos 1,01 ao invés de 1,02 para o intercepto de y. Ambas as respostas são satisfatórias, pois as aplicações não requerem extrema precisão.

(e) $s_e = \sqrt{17.740 - 1,02(584) - 0,734(22.755)} = 4,75$

(g) Somente um ponto cai fora: dessa forma $\dfrac{19}{20} = 0,95$ dentro.

(h) Usando a fórmula (8)

$$t = \frac{0,734 - 0}{4,75}\sqrt{448,8} = 3,3;$$ assim rejeita-se H_0: $\beta = 0$

porque o valor crítico para $v = 18$ é 1,734.

(i) Usando a fórmula (9) com 2,57 substituindo pelo valor de t 2,10 teremos

$$0,734 - 2,10\frac{4,75}{\sqrt{448,8}} < \beta < 0,734 + 2,10\frac{4,75}{\sqrt{448,8}}$$

ou

$$0,263 < \beta < 1,205$$

EXERCÍCIOS

Seção 1

1 — Para os dados seguintes relativos a altura (**x**) e peso (**y**) de 12 estudantes, (a) trace o diagrama de dispersão, (b) adivinhe o valor de **r**, (c) calcule o valor de **r**.

x	65	73	70	68	66	69	75	70	64	72	65	71
y	124	184	161	164	140	154	210	164	126	172	133	150

2 — Calcule o valor de **r** para os seguintes dados sobre escores de teste de inteligência e notas, após ter traçado antes o diagrama de dispersão e ter tentado adivinhar **r**.

T.I.	295	152	214	171	131	178	225	141	116	173	230
Média	3,4	1,6	1,2	1,0	2,0	1,6	2,0	1,4	1,0	3,6	3,6

195	174	236	198	217	143	135	146	227
1,0	2,8	2,8	1,8	2,0	1,2	2,4	2,2	2,4

3 — Por que o valor do coeficiente de correlação não é afetado se trocarmos as variáveis **x** por **y**?

Seção 2

4 — Como imaginaria ser o valor de **r** para os seguintes pares de variáveis: (a) número de homens/hora de trabalho e número de unidades de produto elaborado em certa indústria; (b) tamanho de uma cidade e índice de criminalidade; (c) custo unitário da produção de um artigo e número de unidades produzidas?

5 — Imagine o valor de **r** para os seguintes pares de variáveis: (a) notas de matemática e de línguas estrangeiras; (b) consumo e preço de manteiga; (c) quantidade de chuva na primavera e temperatura média.

6 — Que interpretação você daria se lhe dissessem que a correlação entre o número de acidentes automobilísticos anuais e a idade do motorista é $\mathbf{r} = -0{,}60$ considerados apenas motoristas com pelo menos um acidente?

7 — Explique por que não seria surpreendente encontrar uma alta correlação entre a densidade do tráfico em Wall Street e a altura da maré em Maine se as observações forem feitas hora por hora, a partir das 6 da manhã até às 22 da noite e a maré alta ocorrer às 7 da manhã. Desenhe um diagrama de dispersão dos valores relativos destas duas variáveis para cada hora do dia, que o ajude na sua explicação.

8 — Para os dados do problema 2, mostre aqueles itens para os quais o escore do teste de inteligência é menor que 150 e maior que 225. Agora calcule o valor de **r** e compare com o valor obtido para o problema 2. O que esta comparação parece indicar?

9 — Dê um exemplo de duas variáveis que, sem dúvida, estão altamente relacionadas mas para as quais o valor de **r** seria pequeno pelo fato de a relação não ser linear.

Seção 3

10 — Teste a hipótese de que $\rho = 0$ se uma amostra de tamanho 30 der $\mathbf{r} = 0{,}32$.

11 — A correlação entre aptidão em matemática e línguas estrangeiras, baseada em testes para medir tais aptidões, está por volta de 0,40. Qual seria o tamanho de uma amostra de estudantes para estarmos certos de que o valor do **r** obtido refutaria a afirmação de que $\rho = 0$?

12 — Teste a hipótese $\mathbf{H}_0 : \rho = 0$ contra $\mathbf{H}_1 : \rho > 0$ se uma amostra de tamanho 35 der $\mathbf{r} = 0{,}30$.

13 — Teste a hipótese de que $\rho = 0$ se uma amostra de tamanho 30 der $\mathbf{r} = -0{,}35$. Teria chegado à mesma conclusão se tivesse testado $\mathbf{H}_0 : \rho = 0$ contra $\mathbf{H}_1 : \rho < 0$?

Seção 5

14 — Trace o gráfico da reta cuja equação é $\mathbf{y} = 0{,}3\mathbf{x} - 1$ para valores de **x** entre 0 e 5.

15 — Trace o gráfico da reta cuja equação é $\mathbf{y} = -0{,}5\mathbf{x} + 1{,}5$ para valores de **x** entre — 2 e 4.

16 — A linha de regressão para se estimar o gasto familiar anual em alimentos por meio da renda anual é dada por $\mathbf{y} = 300 + 0{,}20\mathbf{x}$. (a) Qual o gasto médio para famílias com rendas de \$ 2.000,00 \$ 5.000,00? (b) Por que você hesitaria em usar esta fórmula para rendas de \$ 0?

17 — Ache a equação da linha de regressão para os dados que seguem. Trace o gráfico no diagrama de dispersão.

x	1	2	3	4	5
y	2	5	4	8	6

18 — Os dados seguintes são para o esforço de tensão e dureza do alumínio fundido sob pressão. Encontrar a equação da linha de regressão para estimar o esforço de tensão a partir da dureza, traçando a mesma no diagrama de dispersão.

F.T.(y)	293	349	368	301	340	308	354	313	322	334	377	247
H (x)	53	70	84	55	78	64	71	53	82	67	70	56

19 — Os dados seguintes são as notas obtidas por estudantes num exame de admissão (**x**) e num exame final (**y**). (a) Encontrar a equação da linha de regressão. (b) Traçar a linha no diagrama de dispersão.

x	129	179	347	328	286	256	477	430	327	245	286	326
y	370	361	405	302	496	323	374	332	435	165	375	466

20 — Dê exemplos de pares de variáveis para as quais se espera que a linha de regressão tenha inclinação negativa.

Seção 6

21 — Explique a falácia da regressão na afirmativa de que os ases do esporte parecem decair de produção após terem estabelecido um recorde.

22 — Dê uma ilustração de dados para os quais poderíamos ver facilmente uma falácia de regressão.

23 — Explique a falácia de regressão na afirmativa de que os astros de futebol americano raramente conseguem no 2.º ano da universidade o que conseguiam quando eram mais novos.

Seção 7

24 — Calcule o erro-padrão de estimativa do problema 17.

25 — Calcule o erro-padrão de estimativa do problema 18 e ache a percentagem de erros que a excede. Que percentagem de erros excede duas vezes o erro-padrão de estimativa?

26 — Explique por que o erro-padrão de estimativa calculado no problema 24 e também no problema 25 não seria seguro para predizer os tamanhos dos futuros erros de estimativa.

Seção 8

27 — Dado $\mathbf{n} = 8$, $\mathbf{b} = 2$, $\mathbf{s}_e = 0{,}2$ e $\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2 = 50$, ache limites de confiança de 95 por cento para β.

28 — Ache limites de confiança de 95 por cento para β no problema 18.

29 — Teste a hipótese de que $\beta = 1{,}0$ no problema 17.

30 — Teste a hipótese de que $\beta = 2{,}5$ no problema 18.

31 — Examinando a fórmula (9), determine como você escolheria 10 valores de **x** entre $\mathbf{x} = 0$ e $\mathbf{x} = 10$ para fazer observações sobre **y** de modo a fazer o intervalo de confiança para β tão pequeno quanto possível.

Geral

32 — De posse dos seguintes dados, resolva as partes (a) a (g) para os primeiros exercícios de revisão da seção 11.

x	1	9	6	4	5	7	7	1	2	8	6	3
y	15	75	55	42	33	45	55	17	32	80	48	45

33 — Com os dados que seguem, (a) ache os pontos; (b) ache a equação da reta dos mínimos quadrados; (c) trace a reta sobre o gráfico que contém os pontos; (d) preveja o valor de **y** para $\mathbf{x} = 16{,}5$; (e) calcule os erros de predição; (f) calcule o desvio-padrão de estimativa; (g) trace duas linhas paralelas à reta dos mínimos quadrados formando uma faixa dentro da qual se possa esperar que estejam 68 por cento dos pontos supostos como distribuídos normalmente, e comente o fato; (h) teste a hipótese de $\beta = 0{,}5$; (i) ache limites de confiança de 90 por cento para β.

x	5	6	7	8	9	10	11	12	13	14	15	16	17
y	7,6	9,5	9,3	10,3	11,1	12,1	13,3	12,7	13,0	13,8	14,6	14,6	14,7

34 — Os seguintes pares de números são as notas obtidas por 30 estudantes num teste de aptidão verbal e num teste de aptidão matemática. Aqui **x** representa a nota do teste verbal e **y** a nota do teste de matemática.

x	y
77	90
57	71
93	79
66	62
86	84
62	66
74	81
70	67
68	62
69	75

x	y
80	80
56	60
66	58
74	60
79	71
54	48
73	76
72	66
63	60
83	93

x	y
42	60
92	85
64	80
71	80
87	92
54	57
65	72
70	70
83	71
87	94

(a) Trace o diagrama de dispersão e observe se a relação pode ser tratada como linear.

(b) Calcule o valor de **r**.

(c) Teste a hipótese de $\rho = 0$.

(d) Ache a equação da reta dos mínimos quadrados.

(e) Calcule o erro-padrão de estimativa por meio da fórmula (7).

(f) Trace o gráfico da linha de regressão; construa uma faixa de $2\mathbf{s}_e$ em torno dela; determine a percentagem dos pontos do diagrama de dispersão que cai dentro desta faixa.

35 — Os dados da Tabela XII do Apêndice podem ser usados para construir problemas de correlação e regressão. Por exemplo, poder-se-ia calcular o coeficiente de correlação entre as variáveis $\mathbf{x}_1$ (escore do teste) e $\mathbf{x}_5$ (avaliação do professor), e determinar se ele refuta a hipótese se $\mathbf{H}_0 : \rho = 0$. Vários problemas de regressão são facilmente construídos com estes dados.

TÓPICOS ADICIONAIS

10

A DISTRIBUIÇÃO QUI-QUADRADO

1 — INTRODUÇÃO

Nos problemas dos capítulos precedentes referentes aos testes de duas médias ou duas proporções, as questões foram resolvidas por meio dos métodos da curva normal. Problemas nos quais há mais de duas variáveis de classificação não podem, contudo, ser resolvidos por tais métodos. Suponhamos, por exemplo, que uma empresa queira determinar se há diferenças entre quatro marcas de máquinas calculadoras com relação ao tempo de desempenho para solução de problemas relativos a negócios. Não se pode resolver o problema de modo eficiente comparando-se quatro médias de tempo, tomando-as duas a duas. Ou então, suponhamos que uma cidade deseje determinar se o índice de acidente automobilístico é o mesmo para os cinco dias úteis da semana. Aqui também seria ineficiente a comparação de 2 dias separadamente.

No próximo capítulo apresentaremos um método para testar a igualdade de várias médias. Neste capítulo estudaremos o problema de compararmos várias proporções. O problema de testarmos a igualdade de duas proporções de população, resolvido no Capítulo 8, é um caso especial de problemas a considerar-se aqui. Tais problemas podem ser considerados da seguinte maneira. Há um número finito, simbolizado por k, de ocorrências possíveis num experimento. Essas possíveis ocorrências são representadas por k células ou casas. O experimento é realizado n vezes e os resultados são expressos registrando-se as ocorrências das freqüências observadas, nas casas correspondentes. O problema então é determinar se as freqüências são compatíveis com as freqüências postuladas por uma teoria.

Exemplo 1. Como ilustração, se o experimento consiste no lançamento de um único dado, haverá 6 casas. Os resultados da realização de tal experimento 60 vezes são registrados na primeira linha das casas do Quadro 1. Se o dado por "honesto", cada face tem a possibilidade de 1/6 de aparecer em cada jogada.

QUADRO 1

	1	2	3	4	5	6
o_i	15	7	4	11	6	17
e_i	10	10	10	10	10	10

Portanto, seria de esperar que cada face aparecesse 10 vezes, em média, num experimento desta natureza. Essas freqüências médias chamam-se freqüências esperadas. Aparecem na linha de casas marcadas com e_i. O problema então é decidir se as freqüências observadas do Quadro 1 são compatíveis com as freqüências esperadas. Em termos do Capítulo 8, este é um problema de testar a hipótese

$$H_0: p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$$

O método geral para testar compatibilidade baseia-se no grau de concordância entre freqüências observadas e esperadas. Esta medida, chamada qui-quadrado, é definida pela fórmula

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{1}$$

Aqui O_i e e_i simbolizam as freqüências observadas e esperadas, respectivamente, para a i-ésima célula e k simboliza o número de células. Para o Quadro 1 o valor do x^2 é dado por

$$\chi^2 = \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10}$$

$$= 13{,}6$$

Ora, é claro do exame da fórmula (1) que o valor do χ^2 será 0 (zero) se houver concordância perfeita e tornar-se-á cada vez maor à medida que as freqüências observadas e esperadas se forem afastando uma da outra. Assim valores de χ^2 sempre maiores podem ser considerados com a expressão de uma concordância experimental pobre. Se dispuséssemos de um dado não viciado e o experimento de lançá-lo 60 vezes fosse repetido um grande número de vezes, cada vez sendo calculado o valor do χ^2, seria obtido um conjunto de χ^2 que poderia ser classificado numa tabela de freqüência relativa e num histograma de χ^2. O histograma diria aproximadamente em que percentagens de tais experimentos poder-se-ia esperar obter as várias amplitudes de valores de χ^2. Então seria possível julgar se o valor de $\chi^2 = 13{,}6$ era anormalmente grande, quando comparado com a seqüência de χ^2 obtida de experimentos com um dado não viciado. Se a percentagem de experimentos para os quais $\chi^2 < 13{,}6$ fosse muito pequena, digamos menos de 5 por cento, poder-se-ia julgar que as freqüências observadas não eram compatíveis com as freqüências esperadas para um dado não viciado; assim, concluir-se-ia que o dado era viciado.

2 — A DISTRIBUIÇÃO QUI-QUADRADO

Justamente como no caso de outras distribuições de freqüências, é possível usar métodos matemáticos para chegar à desejada distribuição teórica de freqüência. Como há somente um número limitado de valores possíveis para as freqüências das casas do Quadro 1, há somente um número limitado de valores de χ^2 possíveis. Assim a distribuição teórica de χ^2 deve ser uma distribuição discreta. Visto que uma distribuição discreta com muitos valores possíveis requer a aplicação de cálculos demorados, considerações práticas exigem uma aproximação contínua simples para a distribuição discreta muito semelhante à aproximação normal para a distribuição binomial. Tal distribuição contínua é conhecida como a distribuição de qui-quadrado. É pena que essa distribuição contínua aproximada da distribuição discreta do qui-quadrado viesse também a ser chamada distribuição do qui-quadrado; contudo, não haverá confusão, pois a distribuição contínua é a única sempre usada.

O gráfico da distribuição contínua do qui-quadrado para o problema do dado é visto na Figura 1. O valor experimental de $\chi^2 = 13{,}6$ foi localizado neste gráfico, juntamente com o valor de $\chi^2 = 11{,}1$ que separa a cauda direita de 5 por cento da distribuição. Uma vez que valores grandes de χ^2 correspondem à concordância

experimental pobre, os valores de χ^2 excedendo 11,1 são escolhidos como região crítica do teste. O valor experimental $\chi^2 = 13{,}6$ cai na região crítica; conseqüentemente é rejeitada a hipótese de que o dado seja não viciado.

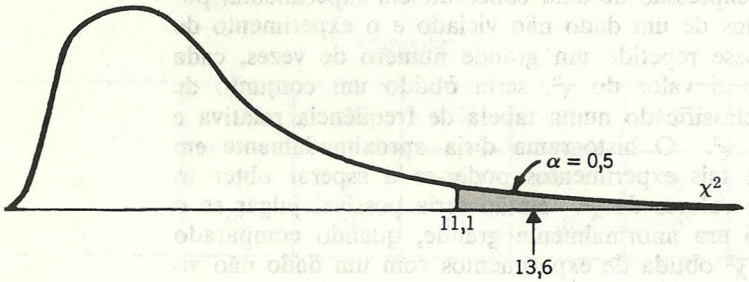


FIGURA 1. Distribuição de χ^2 para problema do dado.

Exemplo 2. Como uma segunda ilustração do teste de χ^2, consideremos o seguinte problema de genética. Em experimentos para obtenção de flores de determinada espécie, um experimentador obteve 120 flores de cor magenta com estigma verde, 48 de cor magenta com estigma vermelho, 36 flores vermelhas com estigma verde e 13 flores vermelhas com estigma vermelho. A teoria mendeliana prevê que flores desses tipos devem ser obtidas nas razões de 9:3:3:1. Esses resultados experimentais são compatíveis com a teoria?

Em termos de notação de testes de hipótese, este é um caso de testarmos a hipótese

$$H_0: p_1 = \frac{9}{16},\quad p_2 = \frac{3}{16},\quad p_3 = \frac{3}{16},\quad p_4 = \frac{1}{16}$$

Para este problema $n = 217$ e $k = 4$. As freqüências esperadas para as quatro casas são obtidas multiplicando-se 217 por cada uma das probabilidades de H_0. As freqüências observadas, juntamente com as freqüências esperadas, após o arredondamento aparecem no Quadro 2.

QUADRO 2

o_i	120	48	36	13
e_i	122	41	41	14

Os cálculos dão

$$\chi^2 = \frac{(120-122)^2}{122} + \frac{(48-41)^2}{41} + \frac{(36-41)^2}{41} + \frac{(13-14)^2}{14} = 1{,}9$$

A teoria da distribuição de χ^2 para problemas tais como esses mostra que a curva de χ^2 na Figura 1 não se aplica ao problema precedente, porque o número de casas neste problema não é o mesmo de antes. Uma característica notável da distribuição de χ^2 é que sua forma depende apenas do número de casas. A Figura 2 mostra o gráfico de seis de tais curvas de χ^2 correspondendo ao número de casas de 2 a 7. É costume classificar-se uma distribuição de χ^2 por meio de um parâmetro $\nu = k - 1$, chamado número de graus de liberdade, ao invés de por meio do número de casas. A expressão *graus de liberdade* refere-se ao número de freqüências de casas independentes. Desde que a soma das quatro freqüências observadas no Quadro 2 deve igualar a 217, a freqüência da quarta célula fica determinada logo que as três primeiras forem especificadas. Assim, há $\nu = 3$ graus de liberdade para este problema, como havia $\nu = 5$ graus de liberdade para o problema anterior do dado. O valor de χ^2 que determina a cauda direita de 5 por cento da distribuição de χ^2

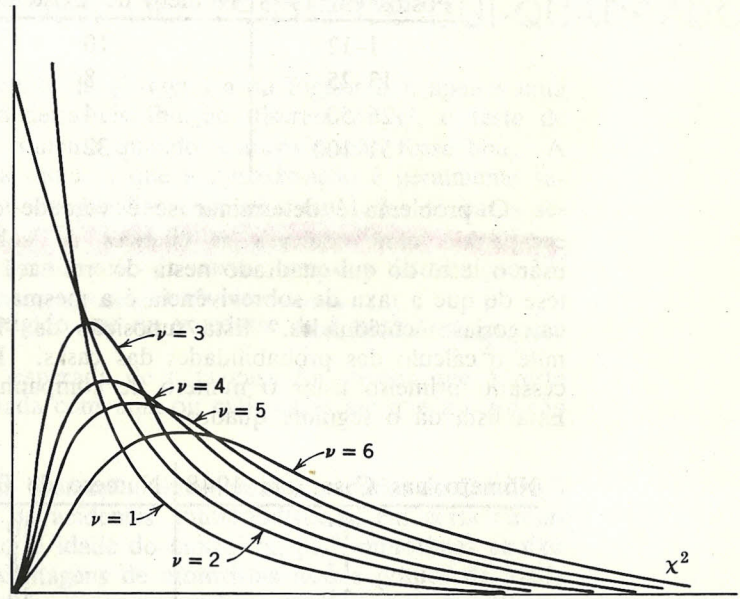


FIGURA 2. Distribuição de χ^2 para vários graus de liberdade.

para $\nu = 3$ verifica-se que é de 7,8. Como o valor de χ^2 para o Quadro 2, a saber, 1,9, não cai na região crítica, o resultado não é significativo. Não há razão com base neste teste para se duvidar da validade da teoria mendeliana, no caso do Quadro 2.

Os valores de χ^2 que determinaram as regiões críticas de 5 por cento nas duas ilustrações precedentes foram obtidos da Tabela VII do Apêndice. O valor crítico de 5 por cento é encontrado no cruzamento da coluna encabeçada por 0,05 com a linha correspondente ao número apropriado de graus de liberdade, $\nu = k - 1$.

O teste de χ^2 pode ser aplicado a problemas para os quais as probabilidades das casas não foram especificadas, mas que podem ser calculadas com base na suposição da "homogeneidade". Segue-se um exemplo desse tipo de problema.

Exemplo 3. Um artigo do *Journal of Industrial Economics,* de novembro de 1972, mostrou as 100 maiores companhias industriais do Reino Unido para o ano de 1948. Dividiu esta lista em 4 grupos: As 12 maiores, as próximas 13 maiores, as seguintes 25 e as restantes 50. Determinou depois quantas dessas companhias estavam ainda vivas no ano de 1968. Os resultados desse estudo aparecem no quadro abaixo.

Postos em 1948	Número de Existentes em 1968
1–12	10
13–25	8
26–50	14
51–100	32

O problema é determinar se é verdade que quanto maior a companhia tanto maiores as chances de sobrevivência. Pode-se usar o teste do qui-quadrado nesta determinação testando-se a hipótese de que a taxa de sobrevivência é a mesma para todas as quatro categorias mencionadas. Esta suposição de "homogeneidade" permite o cálculo das probabilidades das casas. Para fazer isso é necessário primeiro listar o número de companhias nas várias casas. Esta lista dá o seguinte quadro:

Número nas Casas em 1948	Número de Existentes em 1968
12	10
13	8
25	14
50	32

Como o número de companhias originais é de 100, as proporções para essas companhias nas quatro casas para os dados de 1948 são 0,12, 0,13, 0,25 e 0,50. Essas são, portanto, as probabilidades que precisavam ser atribuídas às quatro casas para os dados de 1968, sob a suposição de uma taxa de sobrevivência uniforme. As freqüências esperadas são obtidas multiplicando-se essas probabilidades por 64 que é o número total de companhias sobreviventes. Isso fornece o sequinte quadro de freqüências observadas e esperadas:

o_i	10	8	14	32
e_i	8	8	16	32

O valor de χ^2 é 0,75. Como há 3 graus de liberdade para este problema e o valor crítico é 7,8, aceita-se a hipótese da homogeneidade.

As companhias sobreviventes estão supreendentemente próximas das esperadas na suposição de que o tamanho nada tenha a ver com as chances de sobrevivência.

3 — LIMITAÇÕES DO TESTE DE QUI-QUADRADO

Visto que uma curva de χ^2, como a da Figura 2, é apenas uma aproximação da verdadeira distribuição discreta de χ^2, o teste de χ^2 deveria ser usado somente quando a aproximação fosse boa. A experiência e a teoria indicam que a aproximação é geralmente satisfatória desde que as freqüências esperadas em todas as casas sejam pelo menos iguais a 5. Essa limitação é semelhante à estabelecida para o uso da aproximação da curva normal à distribuição binomial, na qual foi exigido que np excedesse de 5 para $p < \frac{1}{2}$.

Se a freqüência esperada de uma casa for menor que 5, esta casa pode ser combinada com uma ou mais casas até que a condição seja satisfeita.

Exemplo. Como ilustração, consideremos os dados do Quadro 3 sobre a classificação de acidentes automobilísticos, em certa comunidade, de acordo com a idade do motorista, para motoristas abaixo de 25 anos. As percentagens de motoristas nesses grupos de idade na comunidade citada são de aproximadamente 10 por cento, 20 por cento, 20 por cento, 25 por cento e 25 por cento, respectivamente.

O problema aqui é testar se a taxa de acidentes para os motoristas com menos de 25 anos é independente da idade.

Como o número total de observações aqui é $n = 40$, as freqüências esperadas são obtidas multiplicando-se 40 por estas percentagens tratadas como decimais. Foram registradas no Quadro 3 na

QUADRO 3

	15-16	17-18	19-20	21-22	23-24
o_i	5	12	10	8	5
e_i	4	8	8	10	10

linha iniciada por e_i. Porque a freqüência esperada na primeira casa é menor que 5, as duas primeiras casas estão combinadas para dar o Quadro 4.

QUADRO 4

15-18	19-20	21-22	23-24
17	10	8	5
12	8	10	10

O valor de χ^2 para o Quadro 4 é 5,5. Visto que $\nu = 3$ aqui e da Tabela VII o valor crítico de 5 por cento é de 7,8, esse resultado é não significativo.

4 — TABELAS DE CONTINGÊNCIA

Uma aplicação muito útil do teste do χ^2 ocorre em ligação com o teste da comparabilidade das freqüências observadas e esperadas *em tabelas de duas entradas,* conhecidas como *tabelas de contingência.* A Tabela 5, na qual estão registradas as freqüências correspondentes às classificações indicadas para uma amostra de 400, é uma ilustração de uma tabela de contingência.

Uma tabela de contingência usualmente é construída com o propósito de estudar a relação entre as duas variáveis de classificação. Em particular, pode-se desejar saber se as duas variáveis estão relacionadas de algum modo. Por meio do teste do χ^2 é possível testar a hipótese de as duas variáveis serem independentes. Assim, com referência ao Quadro 5, o teste do χ^2 pode ser usado para

QUADRO 5

Educação	Ajustamento Matrimonial				Totais
	Muito baixo	Baixo	Alto	Muito alto	
Faculdade	18 (27)	29 (39)	70 (64)	115 (102)	232
Segundo grau	17 (13)	28 (19)	30 (32)	41 (51)	116
Primeiro grau	11 (6)	10 (9)	11 (14)	20 (23)	52
Totais	46	67	111	176	400

testar a hipótese de não haver relação entre o nível educacional de um indivíduo e o êxito no seu casamento.

Este problema difere dos problemas precedentes pelo fato de não serem conhecidas as probabilidades de uma observação cair nas várias casas. Como resultado, não é possível escrever as freqüências esperadas para as várias casas, como foi o caso nos outros problemas. Esta dificuldade pode ser superada da maneira que segue.

Consideremos experimentos amostrais repetidos desta espécie nos quais 400 pessoas foram classificadas nas categorias convenientes. Se apenas são considerados aqueles experimentos que produzem os mesmos totais marginais, então as freqüências esperadas podem ser obtidas. Como as extremidades estão agora fixadas, a proporção de graduados em tais amostras de 400 será sempre de $232/400 = 0{,}58$. Portanto, se não há relação entre educação e êxito no casamento, seria de esperar que 58 por cento dos 46 indivíduos da categoria de êxito muito baixo fosse de graduados na universidade. Como 58 por cento de 46 é aproximadamente 27, esta é a freqüência esperada para a primeira casa do Quadro 5 na base de independência. As freqüências esperadas das 3 casas restantes da primeira linha são obtidas de maneira semelhante tomando-se 58 por cento dos totais das colunas. As freqüências das casas da segunda e terceira linha são obtidas usando-se 29 por cento e 13 por

cento, respectivamente, correspondentes às proporções 116/400 e 52/400. Estas freqüências esperadas são as registradas entre parênteses no Quadro 5.

O valor de χ^2 para o Quadro 5 usando os números entre parênteses como freqüências esperadas é 20,7. Embora haja 12 casas nesta tabela, não se escolhe $\nu = 11$ graus de liberdade na consideração do valor crítico de χ^2, como aconteceu nos problemas anteriores. O valor correto de ν a ser escolhido aqui é muito diferente, e é determinado por argumentos estritamente matemáticos. Contudo, é possível adquirir-se sensibilidade para a plausibilidade do valor correto através do seguinte tipo de argumento. Suponha que n seja grande mas de valor fixo. Esperar-se-á então que o χ^2 decresça em variabilidade se o número de casas decrescer, porque haverá menores possibilidades para grandes diferenças entre freqüências observadas e esperadas. Observa-se isso na Figura 2 no deslocamento da distribuição para a esquerda quando ν decresce. Além disso, se o número de casas for fixo mas houver restrições quanto às freqüências que ocorrerem nas células, esperar-se-á também que decresça a variabilidade de χ^2. Ora, em experimentos repetidos do tipo precedente as freqüências das casas foram severamente restringidas pela exigência de que possuam as somas das respectivas linhas e as somas das respectivas colunas. Na primeira linha do Quadro 5, por exemplo, as freqüências observadas devem somar 232. Isso significa essencialmente que as freqüências das primeiras 3 casas dessa linha podem assumir o valor que desejarem contanto que a soma não seja maior que 232, mas então a freqüência da quarta célula estará completamente determinada. Assim, se o teste do χ^2 tiver de ser aplicado somente à primeira linha empregaria $\nu = 3$ graus de liberdade. A expressão "graus de liberdade" descreve corretamente o que corresponde ao número de freqüências de casas livres para variarem. O mesmo argumento valeria para a segunda linha dando assim um total de 6 graus de liberdade para as duas linhas. Contudo, quando se olha para a terceira linha, o quadro muda, porque agora é necessário evidenciar que aquilo que vale para linhas deve valer também para colunas; portanto a freqüência encontrada na última casa da primeira coluna é determinada quando as freqüências das duas primeiras casas desta coluna foram especificadas, porque a soma da coluna foi fixada. Como resultado, tanto as freqüências da terceira linha quanto da quarta coluna ficam determinadas ao se especificarem as freqüências das casas restantes. Assim, parece razoável que o experimento deva comportar-se como um experimento no qual há cinco casas nas quais as freqüências podem variar, sujeitas somente à restrição de que suas somas não sejam excessivas. Em vista de argumentos desse tipo não seria surpreendente sabermos que a teoria matemática demonstrou que se deveria escolher $\nu = 6$ aqui.

Da Tabela VII, o valor crítico de 5 por cento de χ^2 com $\nu = 6$ graus de liberdade é 12,6. Como $\chi^2 = 20{,}7$ aqui, este resultado é significativo e rejeita-se a hipótese da independência. Uma inspeção do Quadro 5 mostra que indivíduos com alguma educação colegial parecem ajustar-se melhor ao casamento do que aqueles que não a possuem.

Na solução precedente, como só foram considerados experimentos com o mesmo total marginal supõe-se tacitamente que qualquer relação ou falta de relação existente para experimentos amostrais restritos valerá também para experimentos amostrais irrestritos.

Esse problema ilustra muito bem a razão pela qual a Tabela VI para χ^2 registra o número de graus de liberdade ν de preferência ao número de casas a fim de determinar a curva do χ^2 a usar. Para uma tabela de contingência com r linhas e c colunas, o número de graus de liberdade é dado pela fórmula

$$\nu = (r - 1)(c - 1)$$

Isso decorre do fato de as freqüências na última linha e última coluna ficarem determinadas pelos totais marginais logo que são dadas as freqüências das outras casas. Assim, o número de freqüências de casas independentes é obtido contando-se o número de casas após terem sido suprimidas a última linha e a última coluna. Após a supressão haverá $r - 1$ linhas e $c - 1$ colunas e portanto $(r - 1)(c - 1)$ casas.

Embora a distribuição do χ^2 só dependa do número de casas numa tabela de contingência o valor do χ^2 tende a crescer à medida que cresce o tamanho da amostra ainda que as proporções de linhas e colunas permaneçam as mesmas. Para observarmos o efeito do aumento do tamanho amostral, consideremos o seguinte problema.

Exemplo. Uma amostra de 120 estudantes foi classificada com respeito a aparência e com respeito a notas. Usaram-se três categorias de aparência: atraente, normal e não atraente e quatro categorias de notas. As últimas foram escolhidas da mais alta à mais baixa para fornecer aproximadamente o mesmo número de estudantes em cada categoria. Esta classificação apresentou os seguintes valores das tabelas de contingência mostrados no Quadro 6. O problema é testar a hipótese de que não há relação entre duas variáveis de classificação.

QUADRO 6

	H	H-M	M-L	L
A	14	11	10	5
O	10	16	16	14
U	3	4	7	10

Somando as freqüências das linhas e colunas temos os totais que aparecem no Quadro 7. Tomando a proporção $\frac{40}{120}$ ou $\frac{1}{3}$

QUADRO 7

14 (9)	11 (10)	10 (11)	5 (10)	40
10 (13)	16 (14)	16 (15)	14 (14)	56
3 (5)	4 (6)	7 (7)	10 (6)	24
27	31	33	29	120

do total de cada coluna temos as freqüências esperadas, em parênteses, na primeira linha. As freqüências esperadas para a segunda e a terceira linha foram obtidas de maneira semelhante tomando-se as proporções 56/120 e 24/120 dos totais das colunas. Os cálculos aqui foram arredondados.

O cálculo do χ^2 então é:

$$\chi^2 = \frac{(14-9)^2}{9} + \frac{(11-10)^2}{10} + \frac{(10-11)^2}{11} + \frac{(5-10)^2}{10}$$
$$+ \frac{(10-13)^2}{13} + \frac{(16-14)^2}{14} + \frac{(16-15)^2}{15} + \frac{(14-14)^2}{14}$$
$$+ \frac{(3-5)^2}{5} + \frac{(4-6)^2}{6} + \frac{(7-7)^2}{7} + \frac{(10-6)^2}{6} = 10{,}7$$

Como $\chi_0^2 = 12{,}59$ e $\chi^2 < 12{,}59$, aceita-se a hipótese da não-relação.

Suponha agora que se dobre cada freqüência. As freqüências esperadas dobrarão também. Cada termo do χ^2 será portanto multiplicado por 2. Vê-se isto considerando o primeiro termo que tornará

$$\frac{(28-18)^2}{18} = \frac{[2(14-9)]^2}{2(9)} = 4\frac{(14-9)^2}{2(9)} = 2\frac{(14-9)^2}{9}$$

Como resultado $\chi^2 = 21{,}4$. Como os graus de liberdade não mudaram, $\chi_0^2 = 12{,}59$. Agora rejeita-se a hipótese. Assim, mesmo que as proporções marginais permaneçam as mesmas, o valor do χ^2 tende a crescer à medida que aumenta o tamanho da amostra, desde que as duas variáveis não sejam independentes. Dessa forma, quanto maior for o valor de n tanto mais facilmente se descobre uma falta de independência.

EXERCÍCIOS

Seção 2

1 — Conforme a herança mendeliana, a descendência de certo cruzamento deveria ser vermelha, preta ou branca na seguinte proporção: 9:3:4. Se um experimento mostrou 74, 32 e 38 descendentes nessas categorias, a teoria está confirmada?

2 — O número de indivíduos de certa raça que possuem os 4 tipos de sangue estariam nas proporções de 0,16, 0,48, 0,20 e 0,16. Dadas as freqüências observadas de 180, 360, 130 e 100 para uma outra raça, teste para ver se ela possui a mesma distribuição quanto aos tipos de sangue.

3 — Numa certa cidade a proporção de possuidores de carros com zero, um e mais de um acidentes anuais são respectivamente 0,75, 0,20 e 0,05. Uma companhia de seguros tomou uma amostra de 100 de seus segurados e achou 65, 22 e 13 nessas categorias. Essas freqüências são compatíveis com as proporções teóricas pedidas?

4 — No problema anterior combine as freqüências das duas últimas casas e teste-as. Comente as diferenças dos dois testes.

5 — Tirou-se uma amostra de 100 da distribuição dada pela Figura 5, Capítulo 6, e obtiveram-se as seguintes freqüências correspondentes aos 6 valores possíveis de **x**: 22, 28, 17, 18, 7, 8. Estas freqüências foram obtidas pelo uso da Tabela XI como ficou indicado no experimento amostral deste capítulo. Verifique se esta amostra é compatível com as probabilidades da Figura 5, Capítulo 6.

6 — O número de acidentes automobilísticos semanais de determinada comunidade foram os seguintes: 11, 8, 17, 5, 13, 9, 17, 6, 10, 4. Tais

freqüências estão de acordo com a crença de que as condições dos acidentes são as mesmas, neste período de 10 semanas?

7 — Tome uma amostra de 500 dígitos aleatórios da Tabela XI e liste as freqüências observadas em 10 casas. (a) Use o teste de χ^2 para ver se a amostra é compatível com a teoria aqui. (b) Combine as casas de dígitos ímpares com as de dígitos pares para obter só duas casas e teste para ver se os dígitos ímpares e pares possuem a mesma probabilidade de ocorrência. (c) Poderíamos resolver a parte (b) por uma técnica anterior? Em caso positivo, vamos resolvê-la e comparar os dois resultados.

Seção 3

8 — Certo tablado de jogo tem seis furos de tamanhos variados nos quais a bola pode cair. Os pagamentos pela quadra da bola em dado buraco indica que as probabilidades de sucesso para os vários buracos deveriam ser $\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}$, e $\frac{1}{21}$. Um apostador joga 42 vezes e obtém as seguintes freqüências: 19, 10, 6, 5, 2 e 0. Teste a hipótese de as probabilidades correspondentes aos pagamentos estarem corretas.

9 — Suponhamos que se deseje testar a hipótese de que os pagamentos para os buracos mais difíceis são baixos demais comparados aos pagamentos dos buracos mais fáceis no jogo do problema 8. Um método de se fazer isso é combinar as freqüências dos dois primeiros buracos e dos últimos quatro buracos e testar a hipótese $\mathbf{H_0}$: $\mathbf{p} = \frac{11}{21}$, onde **p** é a probabilidade de queda em um dos dois primeiros buracos. Efetue este teste.

10 — Os dados seguintes representam os resultados de uma investigação da distribuição de sexo das crianças de 32 famílias possuindo cada uma delas 4 crianças. Use a distribuição binomial com $\mathbf{n} = 4$ e $\mathbf{p} = \frac{1}{2}$ para calcular as freqüências esperadas. Aplique depois o teste de χ^2 para ver se este modelo de distribuição binomial é satisfatório aqui.

Número de filhos	0	1	2	3	4
Número de famílias	4	10	8	7	3

Seção 4

11 — Teste para ver se as duas variáveis de classificação da seguinte tabela de contingência são independentes.

20	10	10
10	20	30

12 — Afirma-se que certa droga é eficiente na curva de resfriados. Numa experiência com 164 pessoas resfriadas a droga foi dada à metade delas. A outra metade recebeu pílulas de açúcar. As reações dos pacientes ao tratamento aparecem no quadro abaixo. Teste a hipótese de que as pílulas de açúcar e a droga causam as mesmas reações.

	Ajudou	*Prejudicou*	*Não teve efeito*
Droga	50	10	22
Açúcar	44	12	26

13 — Um analista de mercado está preocupado em saber se as opiniões das donas de casa que afirmam não estar em casa quando procuradas por um analista diferem das opiniões das outras donas de casa, com relação a certo produto. Para verificar esta possibilidade, os entrevistadores voltavam a procurar pelas donas de casa quantas vezes fosse necessário até conseguirem uma entrevista. Os resultados de tais estudos aparecem no quadro abaixo. Procure verificar através de um teste se as mulheres que "não estão em casa" têm a mesma opinião das mulheres que "estão em casa".

Opinião de Produto	*Número de Donas de Casa Entrevistadas*	
	Primeira Chamada	*Última Chamada*
Excelente	62	34
Satisfatório	84	44
Insatisfatório	24	22

14 — Numa hipotética epidemia 927 crianças contraíram a doença. Destas, 408 não receberam tratamento e 104 delas sofreram efeitos posteriores. Das restantes que efetivamente receberam tratamento, 166 tiveram recaída. Teste a hipótese de que o tratamento não foi efetivo e comente esta conclusão.

15 — Resolva o problema 14 através do método explicado no Capítulo 8 para testar a diferença de duas proporções. Pode-se demonstrar que os dois métodos são equivalentes para problemas deste tipo.

16 — O quadro seguinte está baseado numa pesquisa sobre médicos de certa cidade. Foram divididos em três grupos com base em suas notas na escola de medicina e na renda após 15 anos de prática médica. Procure ver através de teste se há alguma relação entre essas duas variáveis.

		Renda		
		Alta	*Média*	*Baixa*
Registro Acadêmico	Alto	18	17	5
	Médio	26	38	16
	Baixo	6	15	9

Geral

17 — Fez-se um estudo para determinar se os motoristas mais antigos, especialmente aqueles com problemas médicos, possuíam maior probabilidade de sofrer acidentes do que os jovens. Os resultados deste estudo foram relatados no **Journal of Chronic Diseases**, de julho-dezembro de 1967. O Quadro abaixo dá os resultados de parte destes estudos.

Categoria Médica	*Número de Sujeitos*	*Número de Acidentes*
Sãos, 30-59	267	22
Sãos, $\geqslant 60$	83	7
Senilidade, $\geqslant 60$	82	8
Distúrbios cárdio-vasculares ($\geqslant 60$)	80	6
Distúrbios cárdio-vasculares + senilidade ($\geqslant 60$)	199	33
Totais	711	76

(a) Calcule o número esperado de acidentes para cada categoria médica, supondo ser constante o índice de acidentes nessas categorias. Aplique depois o teste de χ^2 para ver se esta suposição se justifica.

(b) Elimine o último grupo, combine os grupos restantes $\geqslant 60$ e teste para ver se o índice de acidentes é o mesmo que o do grupo de idade de 30-59.

(c) Quais são as implicações práticas deste estudo com relação aos prêmios que as seguradoras devem impor aos motoristas mais antigos? As seguradoras não vendem apólices a indivíduos com mais de 65 anos de idade.

18 — A **Harvard Business Review**, de julho-agosto de 1973, relatou um estudo feito para determinar se havia uniformidade nos serviços de saúde de oito grandes sistemas privados de saúde. Um dos aspectos deste estudo foi determinar se a taxa de admissão nos hospitais era a mesma para todos os sistemas. O quadro seguinte fornece o número de admissões para cada sistema, num determinado período. Tal quadro foi elaborado a partir de uma amostra aleatória de 1 000 membros de cada sistema. (a) Aplique o teste de χ^2 para ver se se pode supor constante a taxa de admissão nos oito sistemas. (b) Combine os sete primeiros grupos e teste a homogeneidade com relação à Cruz Azul.

	Plano de Saúde							
	1	2	3	4	5	6	7	*Cruz Azul*
Admissões	80	94	89	84	91	81	64	123

19 — Aplique um teste de χ^2 aos dados dos fumantes de cigarro do problema 30, Capítulo 7, para determinar se há alguma relação entre o grau de enfisema e a quantidade de cigarros fumados, para indivíduos que fumam pelo menos um maço por dia. Será preciso combinar algumas das linhas para nos certificarmos de que as freqüências de todas as casas esperadas é, pelo menos, igual a 5.

11

ANÁLISE DE VARIÂNCIA

1 — INTRODUÇÃO

Talvez o problema mais freqüente no trabalho estatístico seja testar se duas amostras diferem significativamente com relação a alguma propriedade. No Capítulo 8 resolveu-se este problema testando a igualdade de duas médias populacionais ou de duas proporções populacionais.

A razão pela qual este tipo de problema ocorre tão freqüentemente é que os experimentadores quase sempre projetam um experimento para comparar uma nova técnica ou processo com uma técnica ou processo-padrão. Por exemplo, um educador pode julgar que descobriu uma maneira mais eficiente de ensinar línguas estrangeiras do que a usada até então; ou um químico pode ter descoberto um novo plástico que reputa superior ao usado na sua fábrica. Em ambos os casos, é preciso elaborar um novo experimento para testar se o novo método ou o novo material é realmente superior ao antigo.

Mas, freqüentemente ocorrem situações em que há vários métodos ou produtos e não apenas dois competindo entre si. Por exemplo, um produto de misturas de bolo pode variar a quantidade de certos ingredientes para obter seis diferentes misturas que gostaria de comparar quanto à qualidade. Ora, geralmente é muito incômodo e ineficiente compararmos duas amostras tomando-as duas a duas. Se tivéssemos, digamos, seis amostras para compararmos, haveria quinze de tais pares de comparação. Além disso, a probabilidade associada com o ato de testarmos uma única diferença não é mais aplicável quando tivermos de testar várias diferenças simultaneamente. Outra desvantagem do método baseado só na comparação de

pares de amostras é que os experimentadores só acostumados a este tipo de comparação são levados a criar experimentos pobres para conseguir seus objetivos finais. O produtor de misturas para bolo, por exemplo, que só mudava um único ingrediente por vez e depois retinha a melhor das duas misturas poderia ter-se saído melhor ao tentar uma mistura variando diversos ingredientes e alterando suas quantidades e considerando, ao mesmo tempo, várias combinações de misturas. Em experiências agrícolas relacionadas ao teste de diferentes tipos e quantidades de fertilizantes e a diferentes variedades de sementes perde-se muito em eficiência se não se consideram as diferentes combinações dessas quantidades simultaneamente.

Em vista da discussão precedente, parece claro que precisamos de um novo método para resolvermos alguns dos problemas relacionados a várias variáveis. Um dos métodos destinados a resolver problemas desse tipo para variáveis contínuas é conhecido como *análise de variância*. É costume abreviar esta expressão escrevendo-a ANOVA. Como o nome indica, o método consiste em dividir a variância da amostra em componentes úteis. Embora o método tenha sido resolvido para tratar ampla gama de problemas, só consideraremos neste capítulo duas dessas aplicações.

2 — UMA VARIÁVEL DE CLASSIFICAÇÃO

O tipo mais simples de modelo de análise de variância é aquele no qual as observações são classificadas em grupos baseados em uma única propriedade. Os valores observacionais de cada grupo são registrados em colunas separadas da tabela. Haverá tantas entradas na coluna quantas forem as experiências.

Com o propósito de explicar a técnica da análise de variância, consideremos os dados do Quadro 1. Estes dados representam os pontos conseguidos por 24 datilógrafos durante um experimento para determinar se há diferenças entre três diferentes marcas de máquina de escrever. Os 24 datilógrafos foram divididos em 3 grupos iguais por escolha casual. Cada grupo de 8 foi então designado para um tipo de máquina. Se não houver desvantagem em uma marca de máquina sobre qualquer das outras e se os pontos da média da população correspondente às três marcas forem simbolizadas por μ_1, μ_2 e μ_3, o problema se reduzirá ao de testar esta hipótese:

(1) $$H_0: \mu_1 = \mu_2 = \mu_3$$

QUADRO 1

I	II	III
44	40	54
39	37	50
33	28	40
56	53	55
43	38	45
56	51	66
47	45	49
58	60	65

Quando a hipótese H_0 é verdadeira, a classificação dos dados em 3 colunas não tem sentido, e o conjunto total de medidas pode ser tratado como uma amostra de tamanho 24 de uma única população. Isto significa que a variação de pontos ao datilografar é devida ao fato de o datilógrafo não ter nada a ver com as diferenças de máquina. Portanto, se a hipótese for verdadeira, supõe-se que o experimento é equivalente àquele no qual cada um dos 24 datilógrafos usa a mesma máquina. Se σ^2 denota a variância da população, pode-se obter uma estimativa de σ^2 por meio da variância amostral familiar baseada nessas 24 medidas. Contudo, há várias outras maneiras de conseguirmos estimativas válidas de σ^2. Por exemplo, a variância da amostra da primeira coluna de medidas é uma estimativa válida não tendenciosa de σ^2, embora não seja tão boa como um estimativa baseada em todas as medidas. Do mesmo modo, as variâncias amostrais da segunda e da terceira coluna são também estimativas válidas de σ^2. Além disso, a média das estimativas de σ^2 das três colunas é uma estimativa válida e não tendenciosa de σ^2 e quase tão boa quanto a estimativa familiar baseada na combinação de três conjuntos de medidas. Se $s^2{}_1$, $s^2{}_2$ e $s^2{}_3$ significam as variâncias amostrais para as três colunas, esta última estimativa, simbolizada por V_c, é

$$V_c = \frac{s_1{}^2 + s_2{}^2 + s_3{}^2}{3} \tag{2}$$

A letra c minúscula é usada aqui para indicar que a estimativa está baseada nas variâncias da coluna. Outro tipo muito diferente de

estimativa de σ^2 pode-se obter usando a relação entre a variância de uma média amostral e a variância da população, isto é $\sigma^2_{\bar{x}} = \sigma^2/n$.

É conveniente aqui expressar esta relação na forma

$$\sigma^2 = n\sigma_{\bar{x}}^2 \tag{3}$$

Suponhamos que várias amostras cada uma de tamanho n foram tiradas de alguma população. Se as médias das amostras tiverem sido calculadas, a variância amostral dessas médias amostrais será uma estimativa válida de $\sigma^2_{\bar{x}}$. Em geral, a variância amostral de um conjunto de medidas é uma estimativa válida de variância da população das medidas quer estas medidas sejam medidas simples ou médias de medidas simples ou outras funções de medidas simples. Da fórmula (3) segue-se que se se dispõe de uma estimativa de $\sigma^2_{\bar{x}}$ devemos multiplicá-la por n para obtermos uma estimativa de σ^2. No presente problema há três destas médias amostrais que podem ser usadas para construir uma estimativa de $\sigma^2_{\bar{x}}$. São as médias das três colunas, simbolizadas por $\bar{x}_1$, $\bar{x}_2$ e $\bar{x}_3$. Ora, a média dessas médias das três colunas é igual à grande média $\bar{x}$, de todas as medidas; dessa forma, uma estimativa não tendenciosa de $\sigma_{\bar{x}}^2$ é dada por $\sum_{j=1}^{3} (\bar{x}_j - \bar{x})^2/2$. Como estas médias estão baseadas em amostras de tamanho 8 cada uma segue-se que $n = 8$ e, assim, da fórmula (3) que a estimativa desejada de σ^2 é dada por

$$V_m = 8\frac{\sum_{j=1}^{3}(\bar{x}_j - \bar{x})^2}{2} \tag{4}$$

A letra m minúscula é usada aqui para indicar que a estimativa está baseada nas médias das colunas.

Como V_c e V_m são ambas as estimativas válidas não tendenciosas de σ^2 quando H_0 é verdadeira, segue-se que deveriam ter valores aproximadamente iguais e, portanto, sua razão deveria estar próxima de 1. Se, contudo, H_0 não for verdadeira e as médias das colunas diferirem consideravelmente, as duas estimativas V_c e V_m também diferirão muito em valor. Porque a estimativa V_c está baseada no cálculo das variâncias de cada coluna separadamente, não será afetada pela alteração das médias das várias colunas, pois a variância de um conjunto de medidas é independente do valor de sua média. É claro da fórmula (4), contudo, que a estimativa V_m será direta-

mente afetada e crescerá em valor à medida que as médias amostrais se afastarem uma da outra.

Assim, parece que a razão V_m/V_c excederá 1 quando H_0 não for verdadeira. Esta razão será usada como a quantidade desejada para testarmos a hipótese H_0. É simbolizada pela letra F; dessa forma

$$F = \frac{V_m}{V_c} \tag{5}$$

Na experiência relativa às máquinas, os cálculos com os dados do Quadro 1 forneceram os valores $\bar{x}_1 = 47$, $\bar{x}_2 = 44$, $\bar{x}_3 = 53$, $s_1^2 = 81{,}1$, $s_2^2 = 106{,}3$ e $s_3^2 = 82{,}3$. Como resultado, segue-se da fórmula (2) que $V_c = 89{,}9$. Cálculos adicionais fornecem o valor $\sum_{j=1}^{3} (\bar{x}_j - \bar{x})^2/2 = 21$. Dessa forma, segue-se da fórmula (4) que $V_{\bar{m}} = 168$. O valor de F é dado, portanto, por

$$F = \frac{V_m}{V_c} = \frac{168}{89{,}9} = 1{,}87$$

Surge agora a questão: o valor de F é grande demais quando comparado aos valores de F que poderiam ser esperados em experimentos repetidos deste tipo com máquinas idênticas, isto é, na suposição de que H_0 seja verdadeira? Para responder a esta questão é preciso determinar a distribuição amostral de F, o que será feito na próxima seção.

3 — A DISTRIBUIÇÃO DE F

Até aqui nos limitamos principalmente a uma discussão do tipo qualitativo, ao afirmarmos que se pode esperar um valor de F próximo de 1 quando H_0 for verdadeira, e um valor muito maior que 1 quando H_0 não for verdadeira e a média populacional diferir muito. Esta informação não é suficiente para se construir um teste baseado em probabilidade. É necessário saber o que é a distribuição F antes que tais testes sejam realizados.

Como no caso de outras distribuições amostrais, é possível aproximarmos a distribuição de F efetuando repetidos experimentos amostrais do tipo considerado aqui e construir o histograma dos valores de F resultantes. Contudo, a distribuição amostral exata de F pode

ser obtida por métodos matemáticos, desde que se façam as suposições adequadas. As suposições necessárias aqui são que as variáveis das 24 células sejam variáveis normais independentes, todas tendo a mesma média μ e a mesma variância σ^2.

Segue-se que a distribuição de F só depende de quantos dados sejam disponíveis para a estimativa do numerador de σ^2 e quantos sejam disponíveis para a estimativa do denominador. A Tabela IX do Apêndice lista os valores críticos de 5 por cento e de 1 por cento da cauda direita de F correspondente aos diferentes valores dos parâmetros ν_1 e ν_2 chamados número de graus de liberdade do numerador e do denominador de F.

Os graus de liberdade aqui são aqueles que se poderia naturalmente associar à variância amostral usada. Como o número de graus de liberdade para a estimativa usual de σ^2 é dado por $\nu = n - 1$ ou *um* menos que o número de medidas, o número de graus de liberdade para o numerador de F neste problema é dado por $\nu_1 = 2$, porque a estimativa está baseada na média das três amostras. O número de graus de liberdade para o denominador de F neste problema é $\nu_2 = 21$, porque a variância de cada coluna contribui com 7 graus de liberdade e são usadas as variâncias das três colunas.

Na Tabela IX acharemos que o valor crítico de 5 por cento de F correspondente a $\nu_1 = 2$ e $\nu_2 = 21$ é 3,47. Como $F = 1{,}87$ para este problema, aceita-se a hipótese. Os dados estão de acordo com o ponto de vista de que a habilidade de datilografar não é afetada pelas três marcas das máquinas usadas.

Embora a abordagem usada para se chegar à variável F no sentido de testar a hipótese de que um conjunto de médias da coluna sejam iguais pareça muito diferente da usada no Capítulo 8 para testar a mesma igualdade, pode-se demonstrar que o teste F, quando aplicado para testar a igualdade das médias das duas colunas é equivalente ao do teste t para o mesmo problema. Assim, o teste baseado em F é uma generalização do teste anterior de duas colunas baseado em t.

4 — NOTAÇÃO ANOVA

O problema anterior é um caso especial de problemas mais gerais deste tipo no qual se tem, digamos, r linhas e c colunas de dados e no qual se deseja testar a hipótese segundo a qual as médias das populações das colunas são iguais. Com a finalidade de considerar tais problemas é conveniente introduzirmos a seguinte notação. Seja

x_{ij} a i-ésima medida da j-ésima coluna do quadro de medidas do qual o Quadro 1 é um caso especial. Seja r o número de linhas e c o número de colunas dessa tabela. A notação aparece no Quadro 2.

QUADRO 2

x_{11}	$x_{12} \cdots x_{1j} \cdots x_{1c}$	$\bar{x}_{1\cdot}$
x_{21}	$x_{22} \cdots x_{2j} \cdots x_{2c}$	$\bar{x}_{2\cdot}$
.	. . .	.
.	. . .	.
.	. . .	.
x_{i1}	$x_{i2} \cdots x_{ij} \cdots x_{ic}$	$\bar{x}_{i\cdot}$
.	. . .	.
.	. . .	.
.	. . .	.
x_{r1}	$x_{r2} \cdots x_{rj} \cdots x_{rc}$	$\bar{x}_{r\cdot}$
$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2} \cdots \bar{x}_{\cdot j} \cdots \bar{x}_{\cdot c}$	$\bar{x}$

A média das medidas da j-ésima coluna é denotada pelo símbolo $\bar{x}_j$. Coloca-se um ponto na frane de j para indicar que a média foi obtida da soma do índice i (linhas). Na próxima seção será necessário somar as colunas também; portanto, é preciso uma notação desse tipo para se saber se estamos somando linhas ou colunas.
Em termos desta nova notação, a variância das médias da coluna é dada por

$$\frac{\sum_{j=1}^{c} (\bar{x}_{\cdot j} - \bar{x})^2}{c - 1}$$

Como cada média da coluna baseia-se em medidas de r, é necessário multiplicar a variância por r para se conseguir uma estimativa da variância σ^2 da população. Esta corresponde à fórmula (3) com n substituído por r. Portanto, o numerador de F é dado pela fórmula

$$V_m = r\frac{\sum_{j=1}^{c} (\bar{x}_{\cdot j} - \bar{x})^2}{c - 1} \tag{6}$$

Para obtermos uma fórmula para V_c, é preciso calcular a variância das medidas da j-ésima coluna. Esta variância, simbolizada por s_j^2, é dada por

$$s_j^2 = \frac{\sum_{i=1}^{r} (x_{ij} - \bar{x}_{\cdot j})^2}{r - 1}$$

A estimativa desejada de σ^2 que deve ser usada no denominador de F é a média das variâncias dessas colunas, que é

$$V_c = \frac{s_1^2 + s_2^2 + \cdots + s_c^2}{c}$$

Em termos de somas de quadrados, segue-se portanto que

$$V_c = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x}_{\cdot j})^2}{c(r - 1)} \tag{7}$$

O quociente das fórmulas (6) e (7) fornece a variável F desejada, com os graus de liberdade dados por $\nu_1 = c - 1$ e $\nu_2 = c\,(r-1)$; assim, a fórmula geral desejada para F é dada por

$$F = \frac{rc(r - 1)}{c - 1}\frac{\sum_{j=1}^{c} (\bar{x}_{\cdot j} - \bar{x})^2}{\sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x}_{\cdot j})^2} \tag{8}$$

Ao se efetuar uma investigação ANOVA é costume apresentar a informação essencial necessária para o teste em forma de tabela. Como esta forma de tabela permaneceu inalterada durante anos e portanto é facilmente inteligível por indivíduos interessados em experimentos ANOVA feitos por outros, ela permite que o leitor de um artigo de pesquisa que usa ANOVA entenda a situação dessa pesquisa com muita rapidez.

Para simplificarmos uma tabela ANOVA, é conveniente introduzir a seguinte notação para as várias somas de quadrados que ocorrem na análise.

$$S_c = \sum_{i=1}^{r}\sum_{j=1}^{c} (\bar{x}_{.j} - \bar{x})^2 = r\sum_{j=1}^{c} (\bar{x}_{.j} - \bar{x})^2$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x}_{.j})^2$$

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{c} (x_{ij} - \bar{x})^2$$

O primeiro símbolo, S_c, expressa a soma dos quadrados das *médias da coluna*. O segundo símbolo, S_e, expressa a soma dos quadrados do *erro experimental*. Mede a variação dos dados independente de quaisquer diferenças de médias de colunas, e assim representa a variação natural ou experimental de X_{ij} quando H_0 é verdadeira. O terceiro símbolo, S_T expressa a soma *total* dos quadrados. Pode-se demonstrar por meio de álgebra que $S_c + S_e = S_T$. Não é preciso aqui a soma total dos quadrados, mas teremos de a utilizar na próxima seção. Em termos dessa notação, a tabela ANOVA assume a forma mostrada no Quadro 3.

QUADRO 3

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado da média	F
Médias da coluna	S_c	$c - 1$	$\frac{S_c}{c-1}$	$\frac{S_c}{c-1} \div \frac{S_e}{c(r-1)}$
Erro experimental	S_e	$c(r-1)$	$\frac{S_e}{c(r-1)}$	
Totais	S_T	$cr - 1$		

A tabela ANOVA particular para o experimento das máquinas de escrever aparece no Quadro 4.

QUADRO 4

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado da média	F
Médias da coluna	336	2	168	$\frac{168}{89{,}9} = 1{,}87$
Erro experimental	1888	21	89,9	
Totais		23		

5 — DUAS VARIÁVEIS DE CLASSIFICAÇÃO

O problema precedente de análise de variância foi relativamente simples porque só havia uma variável de classificação, isto é, marca da máquina. Num experimento deste tipo poder-se-ia considerar muitas outras variáveis que poderiam influenciar os escores. Por exemplo, poder-se-ia considerar as diversas experiências dos datilógrafos e o conhecimento deles com relação à marca das máquinas que lhes foram designadas. Os métodos de análise de variância foram elaborados para tratar qualquer número de variáveis de classificação; contudo, a discussão aqui limita-se a duas dessas variáveis. Os métodos para mais de duas variáveis são muito semelhantes aos de duas variáveis.

O problema considerado na seção precedente pode ser modificado para fornecer um problema que normalmente seria resolvido pelos métodos desta seção. Suponha que, ao invés de termos retirado uma amostra aleatória dos 24 datilógrafos disponíveis, só tivéssemos retirado uma amostra aleatória de 8. Teria sido então necessário para cada um dos 8 datilógrafos familiarizar-se com cada uma das três marcas de máquinas. Depois disso, cada datilógrafo teria sido instruído a datilografar a mesma quantidade de material com cada uma das três marcas. Assim os três escores da primeira linha do Quadro 1 representariam os três escores conseguidos pelo primeiro datilógrafo. Num tal experimento teríamos de escolher aleatoriamente a ordem do emprego dos datilógrafos de modo que nenhuma das marcas levaria, quanto à prática, nenhuma vantagem.

A segunda variável de classificação aqui é o indivíduo. Como bem sabemos que existe uma grande diferença de habilidades entre

datilógrafos, pareceria conveniente controlar esta modalidade de variabilidade dos escores de modo a podermos reconhecer quaisquer diferenças causadas pelas diferentes marcas. Grandes diferenças individuais entre os 24 datilógrafos do experimento anterior encobririam quaisquer diferenças moderadas surgidas das diferentes marcas de máquinas.

Na estrutura de análise de variância com duas variáveis supõe-se que cada uma das variáveis X_{ij} do Quadro 2 seja uma variável normal independente com uma variância comum σ^2. Supõe-se também que as variáveis da linha e as variáveis da coluna não mudem com experimentos repetidos. Isso significa, por exemplo, que em experimentos repetidos de datilografia do tipo discutido os mesmos 8 datilógrafos usam cada uma das três marcas, no decorrer do experimento. Isto difere da estrutura de classificação de uma variável onde sempre se deveria escolher conjuntos novos de 24 datilógrafos. Para a situação de duas variáveis precisamos também fazer mais alguns pressupostos sobre as variáveis básicas, mas tais pressupostos não serão discutidos aqui.

Como no método anterior, acham-se duas estimativas da variância σ^2 comum e usa-se então a razão dessas duas estimativas para se obter o valor F. O método de se encontrar tais estimativas baseia-se na tomada da estimativa da variância natural e em dividi-la em componentes úteis. Tal processo deu origem ao nome "análise de variância". A estimativa da variância natural para o Quadro 2 é a quantidade

$$\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x})^2}{rc-1}$$

Só o numerador dessa estimativa é usado na análise que segue. Ora, pode-se demonstrar por simples manipulação algébrica que a fórmula seguinte é válida:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x})^2=\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{i\cdot}-\bar{x})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{\cdot j}-\bar{x})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2 \tag{9}$$

Pode-se também demonstrar que cada uma das três somas dos quadrados do segundo membro da fórmula (9), se dividida pela respectiva constante, é uma estimativa não tendenciosa de σ^2 quando se supõe que não haja diferenças reais nas médias das linhas ou nas médias das colunas. Estas três estimativas são

$$V_r=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{i\cdot}-\bar{x})^2}{r-1}=c\frac{\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2}{r-1}$$

$$V_c=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{\cdot j}-\bar{x})^2}{c-1}=r\frac{\sum_{j=1}^{c}(\bar{x}_{\cdot j}-\bar{x})^2}{c-1}$$

$$V_e=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2}{(r-1)(c-1)}$$

As letras minúsculas sob V referem-se a filas, colunas ou erro experimental. A expressão V_c é exatamente a mesma que V_m dada pela fórmula (6). A expressão V_r é semelhante, só que mede a variação da linha e não a variação da coluna; portanto se houver grandes diferenças de habilidades entre os datilógrafos, esta quantidade tenderá a ser muito maior do que se não houvesse diferenças entre eles. A expressão V_e mede essencialmente a variação dos dados após a eliminação das desigualdades provocadas por diferenças de coluna e diferenças de linha. Serve como uma estimativa de σ^2 não afetada por diferenças de marca ou individuais.

Para testarmos a hipótese que afirma serem iguais todas as médias teóricas das colunas, temos de usar V_c e V_e. Dessa forma, o teste se reduz a calcular o valor de F dado por

$$F=\frac{r(r-1)\sum_{j=1}^{c}(\bar{x}_{\cdot j}-\bar{x})^2}{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2} \tag{10}$$

$$\nu_1=c-1, \qquad \nu_2=(r-1)(c-1)$$

Os valores de ν_1 e ν_2 em F são sempre os denominadores necessários para se fazer as somas correspondentes das estimativas não tendenciosas dos quadrados de σ^2. A teoria matemática relativa a este problema mostra que a distribuição F é válida aqui, havendo ou não havendo diferenças reais nas médias das linhas.

No esquema de duas variáveis é possível também testar a hipótese de que as médias teóricas das linhas são iguais, o que implicaria, por exemplo, que não há diferenças entre os 8 datilógrafos no tocante a habilidades. Aqui se usam as estimativas V_r e V_e, e obtém-se a razão F dada por

$$F = \frac{c(c-1)\sum_{i=1}^{r}(\bar{x}_{i\cdot} - \bar{x})^2}{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2} \tag{11}$$

$$\nu_1 = r - 1, \qquad \nu_2 = (r-1)(c-1)$$

Tendo em vista o fato de que quanto maior o número de fontes de variação susceptíveis de serem controladas num experimento tanto maior a probabilidade de se detectar diferenças experimentais de interesse, o teste F baseado na eliminação de variação provocada por diferenças individuais em datilografia e dado pela fórmula (10) seria um teste mais delicado que o usado na seção 4 e dado pela fórmula (8). Com o propósito de compararmos estes dois testes, consideremos a aplicação da fórmula (10) aos dados do Quadro 1.

A soma dos quadrados do numerador é a mesma do teste precedente; portanto, seu valor numérico precisa ser calculado. Geralmente é mais fácil calcular a soma dos quadrados do denominador por meio da fórmula (9) que calculá-la diretamente da definição. Assim calcula-se o primeiro membro da fórmula (9) como as duas somas dos quadrados do segundo membro e depois obtém-se a soma desejada dos quadrados por subtração. Os cálculos dão

$$\sum_{i=1}^{8}\sum_{j=1}^{3}(\bar{x}_{\cdot j} - \bar{x})^2 = 8(42) = 336$$

A primeira soma dos quadrados do segundo membro da fórmula (9) foi calculada da mesma maneira que a segunda soma dos quadrados. Cálculos para os dados do Quadro 1 deram os seguintes valores:

$$\sum_{i=1}^{8}\sum_{j=1}^{3}(x_{ij} - \bar{x})^2 = 2224$$

e

$$\sum_{j=1}^{3}\sum_{i=1}^{8}(\bar{x}_{i\cdot} - \bar{x})^2 = 1768$$

Conseqüentemente, a fórmula (9) fornece o valor

$$\sum_{i=1}^{8}\sum_{j=1}^{3}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 = 2224 - 1768 - 336 = 120$$

O valor de F como aparece na fórmula (10) torna-se então

$$F = \frac{8 \cdot 7(42)}{120} = 19{,}6$$

$$\nu_1 = 2, \qquad \nu_2 = 14$$

Na Tabela IX podemos observar que o valor crítico de 5 por cento de F correspondente a $\nu_1 = 2$ e $\nu_2 = 14$ é 3,74. Como F = 19,6 é a região crítica rejeita-se a hipótese de H_0.

A conclusão aqui é contrária à tirada para os mesmos dados da seção 3. De fato, os dados do Quadro 1 foram obtidos de um grupo de 8 datilógrafos; conseqüentemente, aqui só é aplicável o segundo método. O primeiro método exigiu que os escores fossem aqueles dos 24 datilógrafos selecionados aleatoriamente. O propósito de se usar os mesmos dados para ambos os métodos foi apontar as semelhanças e diferenças dos dois métodos e acentuar que temos um preço a pagar se quisermos introduzir importantes variáveis de classificação na análise da técnica de variância para termos um teste preciso para avaliação de um conjunto de médias teóricas. Se não tivesse havido diferenças individuais apreciáveis nas habilidades referentes à datilografia, nada teria sido ganho ao se elaborar um experimento para medir e eliminar esta fonte de variação do teste, caso em que teria sido preferível o método da seção 3. A aplicação da fórmula (11) mostra, contudo, que $F = 29{,}5$, $\nu_1 = 7$, $\nu_2 = 14$. Como é de 2,77 o valor crítico de 5 por cento de F, isto significa que há grande variação na média das linhas e por isso grande variação na habilidade individual de datilografar. Com o propósito de construirmos uma tabela ANOVA de duas entradas, precisamos acrescentar alguma notação adicional e mudar a definição de S_e. É a seguinte:

$$S_r = \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{i\cdot} - \bar{x})^2 = c\sum_{i=1}^{r}(\bar{x}_{i\cdot} - \bar{x})^2$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$$

QUADRO 5

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado da média	F
Médias da coluna	S_c	$c - 1$	$\frac{S_c}{c-1}$	$\frac{S_e}{\frac{S_c}{c-1} \div (c-1)(r-1)}$
Médias das filas	S_r	$r - 1$	$\frac{S_r}{r-1}$	$\frac{S_e}{\frac{S_r}{r-1} \div (c-1)(r-1)}$
Erro	S_e	$(c - 1)(r - 1)$	$\frac{S_e}{(c-1)(r-1)}$	
Totais	S_T	$rc - 1$		

Com esta notação, a tabela ANOVA de duas entradas assume a forma mostrada no Quadro 5.

A tabela ANOVA para o problema que acabamos de resolver aparece no Quadro 6.

QUADRO 6

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado da Média	F
Médias das colunas	336	2	168	$\frac{168}{8{,}57} = 19{,}6$
Médias das filas	1 768	7	252,6	$\frac{252{,}6}{8{,}57} = 29{,}5$
Erro	120	14	8,57	
Totais	2 224	23		

6 — FÓRMULAS DE CÁLCULO

Se dispomos de equipamento de cálculo para obter a soma dos quadrados que são necessários para aplicar o teste F, geralmente é melhor usar fórmulas que requeiram que achemos só as somas e as somas dos quadrados das entradas do Quadro 2. Tais fórmulas podem ser obtidas por simples manipulações algébricas. As duas primeiras das fórmulas que seguem são necessárias para a fórmula (8), ao passo que todas as fórmulas com exceção da (13) são necessárias para as fórmulas (10) e (11).

$$(12) \qquad \sum_{j=1}^{c} (\bar{x}_{\cdot j} - \bar{x})^2 = \frac{1}{r^2} \sum_{j=1}^{c} \left(\sum_{i=1}^{r} x_{ij} \right)^2 - \frac{1}{r^2 c} \left(\sum_{j=1}^{c} \sum_{i=1}^{r} x_{ij} \right)^2$$

$$(13) \qquad \sum_{j=1}^{c} \sum_{i=1}^{r} (x_{ij} - \bar{x}_{\cdot j})^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} x_{ij}^2 - \frac{1}{r} \sum_{j=1}^{c} \left(\sum_{i=1}^{r} x_{ij} \right)^2$$

$$(14)\qquad \sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2=\frac{1}{c^2}\sum_{i=1}^{r}\left(\sum_{j=1}^{c}x_{ij}\right)^2-\frac{1}{c^2r}\left(\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\right)^2$$

$$(15)\qquad \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x})^2=\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}^2-\frac{1}{rc}\left(\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\right)^2$$

Para ilustrarmos o uso dessas fórmulas de cálculo, analisaremos um segundo conjunto de dados experimentais por meio da técnica de análise de variância. Os dados aparecem no Quadro 7 e representam as colheitas de batata de quatro áreas, cada uma delas dividida em cinco subáreas. Para cada uma das áreas das 5 subáreas foram destinados cinco fertilizantes diferentes. O problema consiste em verificar se os cinco fertilizantes são igualmente efetivos com relação à colheita média.

QUADRO 7

Pontos	Fertilizante				
	A	B	C	D	E
1	310	353	366	299	367
2	284	293	335	264	314
3	307	306	339	311	377
4	267	308	312	266	342

Os cálculos nos dão as seguintes somatórias:

$$\sum_{i=1}^{4}\sum_{j=1}^{5}x_{ij}=6320 \qquad \sum_{i=1}^{4}\sum_{j=1}^{5}x_{ij}^2=2.018.650$$

Cálculos adicionais darão os valores

$$\sum_{j=1}^{5}\left(\sum_{i=1}^{4}x_{ij}\right)^2=8.039.328 \qquad \sum_{i=1}^{4}\left(\sum_{j=1}^{5}x_{ij}\right)^2=10.017.750$$

Como resultado, as fórmulas (12), (14) e (15) darão

$$\sum_{j=1}^{5}(\bar{x}_{\cdot j}-\bar{x})^2=\frac{8.039.328}{16}-\frac{(6320)^2}{80}=3178$$

$$\sum_{i=1}^{4}(\bar{x}_{i\cdot}-\bar{x})^2=\frac{10.017.750}{25}-\frac{(6320)^2}{100}=1286$$

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x})^2=2.018.650-\frac{(6320)^2}{20}=21.530$$

A soma dos quadrados do denominador da fórmula (10) pode ser obtida por meio da fórmula (9). Assim

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2=21.530-5(1286)-4(3178)$$
$$=2388$$

O valor de F na fórmula (10) é, portanto, dado por

$$F=\frac{4\cdot 3\cdot 3178}{2388}=16{,}0$$

$$\nu_1=4 \qquad \nu_2=12$$

Encontraremos na Tabela IX que o valor crítico de 5 por cento de F correspondente a $\nu_1=4$ e $\nu_2=12$ é 3,26. Como $F=16{,}0$ é a região crítica, rejeita-se a hipótese de que os fertilizantes sejam igualmente efetivos. Os fertilizantes E e C parecem superiores aos demais fertilizantes. Experiências ulteriores e testes adicionais poderão dar informações mais precisas concernentes à superioridade relativa de E e C sobre os demais e determinar ainda a taxa de C com relação a E.

O método apresentado na seção 4 é facilmente generalizado para ser aplicado à situação na qual o número de medidas em várias colunas é o mesmo. É também bastante fácil generalizar os métodos da seção 5 de modo que sejam aplicados à situação na qual se têm várias medidas em cada célula do Quadro 2. Uma generalização mais complicada ocorre quando se estendem estes métodos para a situação na qual há várias variáveis de classificação; contudo, os métodos são bastante semelhantes aos que acabamos de explicar.

EXERCÍCIOS

Seção 2 e 3

1 — Os dados seguintes fornecem os resultados de um processo químico resultante da aplicação de 4 catalisadores diferentes. Use a fórmula (8) para verificar se os produtos resultantes são influenciados pelos catalisadores.

I	II	III	IV
36	35	35	34
33	37	39	31
35	36	37	35
34	35	38	32
32	37	39	34
34	36	38	33

2 — Os dados seguintes expressam as colheitas de trigo em campos experimentais correspondentes a 4 tratamentos sulfúricos diferentes para o controle da ferrugem. Os tratamentos consistiam em pulverização antes da chuva, pulverização após a chuva, pulverização uma vez por semana e nenhuma pulverização. Verifique por meio da fórmula (8) se há diferenças significativas nos resultados submetidos aos vários métodos de pulverização.

	Método de Pulverização			
Pontos	1	2	3	4
1	5,3	4,4	8,4	7,4
2	3,7	5,1	6,0	4,3
3	14,3	5,4	4,9	3,5
4	6,5	12,1	9,5	3,8

3 — Os dados seguintes dão os escores obtidos por 24 soldados num experimento destinado a determinar se a precisão do tiro é afetada pela maneira de se dispor os olhos (a) com o olho direito aberto; (b) com o olho esquerdo aberto; (c) com ambos os olhos abertos. Verifique se a precisão é influenciada por estas três maneiras de se encarar o alvo.

Direita	Esquerda	Ambos
44	40	51
39	37	47
33	28	37
56	53	52
43	38	42
56	51	63
47	45	46
58	60	62

4 — No problema 2 use o teste t de Student para determinar se a pulverização teve algum efeito sobre a produção. Em outras palavras, combine os dados para as primeiras três colunas e trate as medidas como se fossem uma amostra de tamanho 12 tirada de uma população normal. As medidas da última coluna podem ser tratadas como uma amostra de tamanho 4 tirada de uma segunda população normal.

Seção 5

5 — Os dados seguintes representam o número de unidades de produção diária obtidas por 5 homens que empregam 4 diferentes tipos de máquinas. (a) Verifique se a produtividade média é a mesma para os 4 diferentes tipos de máquina. (b) Veja se os 5 homens diferem com respeito à produtividade média. Use as fórmulas (10) e (11) aqui.

Trabalhador	Tipo de Máquina			
	1	2	3	4
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

6 — Resolva o problema 2 pelo método das 2 variáveis para ver se a eliminação da variação devida aos campos afetará o teste.

7 — Com relação aos dados do Quadro 7 do texto, teste para ver se há diferenças de produção devido à natureza das áreas plantadas.

8 — Para os dados do Quadro 7, use a fórmula (8) para testar se os fertilizantes têm a mesma efetividade; isto é, suponha que haja 20 espécies diferentes de áreas plantadas.

9 — Levou-se a efeito um experimento para determinar se a produtividade de um processo químico pode ser aumentada alterando-se a temperatura da relação e a quantidade de catalisadores. Os resultados aparecem no quadro abaixo. As entradas da tabela representam a produtividade sob três temperaturas e 4 catalisadores.

(a) Verifique se a produtividade é afetada pela variação do catalisador.

(b) Verifique se a produtividade é afetada pela variação da temperatura.

Temperatura	Catalisador			
	1	2	3	4
A	53	59	58	50
B	57	65	62	60
C	52	62	54	52

12

TESTES NÃO-PARAMÉTRICOS

1 — INTRODUÇÃO

Com uma exceção, todos os métodos apresentados até aqui para testar hipóteses admitiam que a forma de distribuição da população era conhecida, e devia-se fazer um teste sobre alguma suposição a respeito de um parâmetro desta distribuição. Por exemplo, deduziram-se métodos para testar valores de μ para uma distribuição normal e para testar valores de p para uma distribuição binomial. A única exceção ocorreu no Capítulo 10 onde se usou a distribuição do qui-quadrado para se testar a compatibilidade de um conjunto de freqüências observadas com um conjunto de freqüências esperadas. As freqüências esperadas são muitas vezes obtidas sem um conhecimento da distribuição da variável subjacente básica. Por exemplo, no problema da tabela de contingência não foi necessário saber como se distribuíam as duas variáveis básicas a fim de testar se elas eram independentes.

Freqüentemente surgem situações em que não se tem conhecimento da distribuição da variável em estudo ou em que se conhece definitivamente que ela não é do tipo exigido para a aplicação da teoria desejada. Por exemplo: pode-se saber que a variável possui uma distribuição muito diversa da distribuição normal e ainda pode-se querer testar se a média dessa distribuição tem um valor especificado. Uma vez que os métodos de pequena amostra requerem a suposição da normalidade, o método-padrão não seria aplicável a tal problema quando apenas uma pequena amostra fosse disponível.

Elaboraram-se métodos para tomar o lugar dos testes-padrão quando não forem satisfeitas as suposições requeridas por esses testes.

Chamam-se métodos não-paramétricos, pois eles não testam valores de parâmetros de tipos conhecidos de populações, como no caso dos testes-padrão, dos capítulos precedentes. Visto que os métodos não-paramétricos são mais gerais do que os que exigem suposições adicionais, é de se esperar que não sejam tão bons quanto os métodos-padrão, quando ambos forem aplicáveis. Estes novos métodos deveriam conseqüentemente ser usados apenas quando um método-padrão não fosse apropriado.

Os métodos não-paramétricos apresentados neste capítulo foram escolhidos para resolver alguns dos mesmos tipos de problemas resolvidos anteriormente pelos métodos paramétricos. Há ainda uns poucos testes não-paramétricos, válidos para esses mesmos problemas e também para outros tipos.

2 — TESTE DE MÉDIA

Para problemas não-paramétricos relacionados a varáveis contínuas, a mediana é uma medida de posição mais natural para uma distribuição do que a média. A mediana tem a vantagem da propriedade que afirma que a probabilidade é de $\frac{1}{2}$ de que um valor da amostra exceda a mediana da população, qualquer que seja a natureza da distribuição.

Como resultado, é possível elaborar testes para valores hipotéticos da mediana sem saber como é a distribuição subjacente. A mais simples é a prova dos sinais. Com o fim de descrever este teste, consideremos os seguintes dados obtidos de um teste de resistência à ruptura de telhas de cerâmica fabricadas por um novo processo mais barato: 20, 42, 18, 21, 22, 35, 19, 18, 26, 20, 21, 32, 22, 20, 24. Suponhamos que a experiência com o processo antigo tenha produzido uma mediana de resistência à ruptura de 25. A hipótese natural a ser testada aqui é

$$H_0 : \xi = 25 \qquad \text{contra} \qquad H_1 : \xi < 25$$

Onde ξ simboliza a mediana da distribuição.

O primeiro passo ao aplicarmos a prova de sinais é subtrair a mediana pedida de cada medida observada e depois registrar o sinal da diferença correspondente. Se 25 é subtraído de cada um dos valores observados, obtêm-se os seguintes sinais:

$$- \; + \; - \; - \; - \; + \; - \; - \; + \; - \; - \; + \; - \; - \; -$$

A etapa seguinte é contar o número de sinais +, indicado por x, e o número total de sinais indicado por n. Aqui $x = 4$ e $n = 15$. Se a hipótese H_0 for verdadeira, a probabilidade será de $\frac{1}{2}$ de que um sinal + venha a ser obtido ao ser feita uma observação; conseqüentemente a variável x representa o número de sucessos em n tentativas de um experimento para o qual a probabilidade de sucesso numa tentativa isolada é $p = \frac{1}{2}$. O problema fica reduzido a um problema da distribuição binomial do tipo tratado no Capítulo 8.

Se a mediana de uma distribuição for menor que o valor postulado subtraindo-se o valor postulado de um conjunto de medidas observadas, teremos provavelmente mais diferenças negativas que positivas, dessa forma um valor de x menor do que o esperado sob a hipótese. Como valores pequenos de x favorecem a hipótese alternativa, a região crítica do teste deveria estar na cauda esquerda da distribuição binomial.

Cálculos por meio da fórmula dada pela fórmula (1), Capítulo 5, com $n = 15$ e $p = \frac{1}{2}$, forneceram as seguintes probabilidades:

$$P\{0\} = 0{,}000 \qquad P\{3\} = 0{,}014$$
$$P\{1\} = 0{,}000 \qquad P\{4\} = 0{,}042$$
$$P\{2\} = 0{,}003$$

Quando essas probabilidades são somadas, segue-se que $P\{x \leq 4\} = 0{,}059$. Assim, o valor observado de $x = 4$ estaria na região crítica se fosse selecionada uma região crítica do tamanho $\alpha = 0{,}059$, mas não estaria na região crítica se fosse selecionado um valor menor que α.

Se tivéssemos usado aqui a aproximação normal à distribuição binomial baseada no método geométrico exposto na seção 4 do Capítulo 5, teríamos calculado

$$z = \frac{x - np}{\sqrt{npq}} = \frac{4.5 - 7.5}{\sqrt{15 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = -1{,}55$$

Na Tabela IV do Apêndice, $P\{z < -1{,}55\} = 0{,}06$, que concorda muito bem com o resultado obtido pelo cálculo das probabilidades binomiais necessárias.

O problema que acabou de ser resolvido por métodos não-paramétricos corresponde ao primeiro problema de teste de hipótese considerado no Capítulo 8, que consistiu em testar se a média de uma distribuição normal tinha um valor particular. A próxima seção considera um não-paramétrico análogo ao problema de testar se as médias de duas distribuições normais são iguais.

3 — TESTE DA DIFERENÇA DE DUAS MEDIANAS

Como a mediana está sendo usada para substituir a média como parâmetro de posição nos problemas não-paramétricos, é natural testar a diferença de duas medianas em lugar da diferença de duas médias nas situações não-paramétricas. Esta hipótese pode ser escrita sob a forma

$$H_0: \xi_1 = \xi_2$$

onde ξ_1 e ξ_2 indicam as medianas das duas populações de interesse. O teste não-paramétrico introduzido para resolver este tipo de problema é chamado teste da *soma de pontos*. Para ilustrar como o teste é aplicado, consideremos os seguintes dados sobre o número de tentativas necessárias para um grupo de oito ratos tratados e um grupo de dez ratos não tratados aprenderem certa tarefa

T	24	28	15	47	23	25	53	20		
N.T.	22	12	30	16	26	14	18	21	16	18

As duas amostras são primeiro ordenadas conforme o tamanho crescente. Para os dados anteriores temos a seguinte ordem na qual foram sublinhados os itens para o grupo tratado:

12, 14, 15, 16, 16, 18, 18, 20, 21,
22, 23, 24, 25, 26, 28, 30, 47, 53

Estes valores são então substituídos por números indicando sua ordem de colocação para dar

1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 11, 12, 13, 14, 15, 16, 17, 18

A próxima etapa é somar as posições do grupo menor (aqui o grupo tratado). Se esta soma é indicada por R, segue-se que o valor de R será dado pela soma das posições sublinhadas. Para este problema $R = 97$. A etapa final é determinar se o valor de R cai na região crítica do teste.

A distribuição da amostragem de R, sob a suposição de que as duas distribuições de população sejam idênticas, foi elaborada por métodos matemáticos. Esta distribuição depende, é claro, dos tamanhos das duas amostras, indicados por n_1 e n_2. Aqui n_1 indica o menor dos dois tamanhos da amostra. A Tabela VIII do Apêndice fornece os valores críticos desejados correspondentes aos vários tamanhos de amostra para $n_2 \leq 10$. Para maiores tamanhos de amostras, a distribuição de R pode ser aproximada de modo satisfatório pela distribuição normal conveniente. Esta é a distribuição normal com média e desvio-padrão dados pelas fórmulas

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

Para o problema que está sendo discutido julgou-se que os ratos tratados exigiriam um período de aprendizagem maior do que os ratos não tratados; portanto, a hipótese alternativa aqui é a seguinte:

$$H_1: \xi_1 > \xi_2$$

Sob esta alternativa, R tenderia a ser maior do que sob H_0, porque R é a soma das posições do grupo de medidas ξ_1; conseqüentemente deveríamos escolher a região crítica sob a cauda direita da distribuição. Na tabela VIII vamos encontrar que para $n_1 = 8$ e $n_2 = 10$ a probabilidade de que $R \geq 95$ é de 0,051. Como $R = 97$ cai dentro da região crítica, rejeita-se a hipótese.

Se usássemos a aproximação normal teríamos calculado

$$\mu_R = \frac{8(8 + 10 + 1)}{2} = 76$$

e

$$\sigma_R = \sqrt{\frac{8 \cdot 10(8 + 10 + 1)}{12}} = 11{,}3$$

Então calcularíamos

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{97 - 76}{11{,}3} = 1{,}86$$

Na Tabela IV do Apêndice achamos que $P\{z > 1{,}86\} = 0{,}03$; assim rejeita-se H_0. Poderíamos conseguir uma aproximação um pouco mais exata usando 96,5 ao invés de 97 em z.

Quando $n_1 = n_2$ pode-se escolher qualquer grupo como o melhor para o qual R vai ser calculado. Quando os dois grupos de observações contêm um ou mais valores comuns, ocorrerão empates nos números de ordem. Em tais situações é suficiente dar a cada conjunto de valores iguais observados a posição que seja a média das posições ocupadas por eles. Esta modificação não é necessária para empates que ocorrem no mesmo grupo. Visto que a distribuição de R foi obtida na suposição de que as duas distribuições da população fossem idênticas, dever-se-ia realmente testar a hipótese de que as duas distribuições tenham sido idênticas contra a alternativa de que uma delas tenha sido deslocada para a direita. Duas populações podem ter medianas e ainda diferir de maneira a produzir valores de amostra R que regularmente cairiam na região crítica do teste precedente. Assim, não é rigorosamente correto rejeitar a hipótese $H_0 : \xi_1 = \xi_2$ quando R cai na região crítica, a menos que se esteja preparado para admitir que as duas hipóteses sejam idênticas, exceto possivelmente quanto às suas posições. O teste anterior é freqüentemente chamado *teste de deslizamento,* porque determina se duas distribuições são idênticas contra a possibilidade de que uma delas tenha deslizado relativamente à outra.

É interessante comparar o *teste da soma de postos* deste problema com o teste paramétrico correspondente que seria aplicado aqui se pudéssemos supor que as duas variáveis básicas possuíam distribuições normais independentes. Como os tamanhos da amostra são menores, tal comparação seria feita usando-se o teste t de Student, como ficou explicado na seção 6 do Capítulo 8. O que exige que se suponha variâncias iguais para as variáveis além da suposição de normalidade. Os cálculos baseados na fórmula da seção 6, do Capítulo 8, deram o valor de $t = 2{,}18$. Como há $\nu = 16$ graus de liberdade aqui, a Tabela V mostra que este valor de t está muito próximo do valor de 0,025 listado. Esta probabilidade está próxima do valor de probabilidade de 0,03 obtido pela aplicação da aproximação da distribuição normal a R para este problema; portanto, os dois testes diferem muito pouco no tocante a este problema.

O teste da soma dos postos é conhecido como excelente para testes de deslocamento. Mesmo quando há justificativa na suposição de que as duas variáveis estejam distribuídas normalmente de maneira independente com as mesmas variâncias, o teste da soma dos postos é quase tão bom como o teste t de Student, criado para este tipo de problema, no sentido de produzir pequenos erros do tipo II.

Esta propriedade de ser quase tão bom como o teste t quando este se justifica e de ser um teste válido sob todas as condições faz do teste da soma dos pontos um tipo de teste muito atraente.

4 — COEFICIENTE DE CORRELAÇÃO DE POSTOS

O problema de medir a extensão segundo a qual duas variáveis estão linearmente relacionadas foi considerado no Capítulo 9. Aí foi admitido que ambas as variáveis eram medidas numa escala contínua. Além disso, ao se testar se um coeficiente de correlação de população tem um valor zero, foi necessário supor que as duas variáveis tinham uma distribuição normal. Este pressuposto é um tanto restritivo; portanto, seria desejável termos um teste que estivesse livre desta restrição.

O problema de testar se há uma correlação zero entre duas variáveis já foi visto no Capítulo 10, na seção que falou sobre tabelas de contingência. Aí usamos o teste do χ^2 (qui-quadrado) para verificar se as duas variáveis de classificação eram variáveis independentes. Se duas variáveis são independentes, elas também são, evidentemente, não-correlacionadas. Esse método é muito geral pelo fato de não exigir que as variáveis sejam mensuráveis em escala contínua. Basta podermos estabelecer categorias, ou grupos para as variáveis.

Se as variáveis são capazes de serem medidas em escala contínua, como nos problemas de coeficientes de correlação, pode-se então esperar que o método de teste da tabela de contingência para a independência pode ser melhorado, e este é o caso.

O não-paramétrico análogo ao coeficiente de correlação comum é o *coeficiente de correlação de postos.* Como o nome indica, é apenas o coeficiente de correlação calculado para os postos das variáveis ao invés dos valores numéricos das variáveis. Ora, pode-se demonstrar por métodos puramente algébricos que quando x_i e y_i na fórmula r dada pela fórmula (1), Capítulo 9, são tratados como as posições das medidas correspondentes, a fórmula se reduz, então, a

$$r = 1 - 6\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}$$

Embora o valor de amostra de r obtido por meio desta fórmula seja habitualmente próximo do valor obtido por meio do coeficiente de correlação ordinário baseado em medidas, a teoria anterior sobre a distribuição de r não se aplica aqui. Mas, na suposição de que as duas variáveis sejam independentes, a distribuição de r pode ser obtida por métodos matemáticos sem a exigência de um pressuposto de normalidade. A Tabela X do Apêndice fornece os valores críticos de r de 5 por cento e 1 por cento para se testar a hipótese segundo a qual as duas variáveis são independentes. Esta tabela é para testes unicaudais; portanto, se for usada diretamente está-se testando a hipótese de correlação zero contra a alternativa de uma correlação positiva (ou negativa).

Como em outros testes baseados em postos, é costume substituir as posições empatadas pela média das posições ocupadas pelas medidas iguais.

Como ilustração do uso do coeficiente de correlação de postos no teste de independência de duas variáveis, consideremos os seguintes dados sobre os postos atribuídos, por dois críticos de arte amadores, às pinturas de dez estudantes.

Crítico A	7	4	5	8	2	10	1	9	6	3
Crítico B	2	7	4	3	1	10	5	6	9	8

Tomando as diferenças dessas posições, elevando ao quadrado e somando, teremos o valor $\Sigma\ (x_i - y_i)^2 = 120$. Como resultado

$$r = 1 - \frac{6\ (120)}{10\ (100 - 1)} = 0{,}27$$

Da Tabela X, o valor crítico de r de 5 por cento para $n = 10$ é 0,564. Como $r = 0{,}27$ é pequeno demais para ser significativo, aceita-se a hipótese de independência. Não há evidência aqui de que os dois críticos concordem sobre o que é boa pintura. Como mesmo os críticos profissionais freqüentemente discordam, não é surpreendente que juízes amadores discordem.

EXERCÍCIOS

Seção 2

1 — Os dados seguintes representam as idades dos ladrões presos durante o ano passado em certa cidade. Testar a hipótese de que a idade mediana para os ladrões seja de 23 anos, que é a idade obtida da experiência passada, contra a alternativa de que seja menos que 23. Ao fazer o teste, ignorar os indivíduos de 23 anos. As idades são: 24, 16, 18, 19, 18, 17, 26, 28, 23, 24, 19, 21, 21, 22, 17, 15, 21, 36, 18, 38, 19, 22, 21, 26, 21, 33, 21, 17, 26, 18.

2 — Numa escola primária foram constituídos 17 pares de crianças da primeira série com base na semelhança de inteligência e conhecimentos. Uma criança de cada par aprendeu a ler pelo método I e a outra pelo método II. Após um período de treinamento, as crianças fizeram um teste de leitura com os resultados seguintes:

Método I	65	68	70	63	64	62	73	75	72	78	64	73	79	80	67	74	82
Método II	63	68	68	60	65	66	72	74	73	70	66	70	80	78	63	74	78

Tomando as diferenças de valores e ignorando os valores zero, usar o teste do sinal para determinar se os dois métodos são igualmente eficientes.

3 — Resolva o problema 2 por meio do teste t de Student, após tomar as diferenças conforme explicado no Capítulo 8 e comente os dois casos.

4 — Resolva o problema 36, Capítulo 8, por meio do teste de sinal e compare seu resultado com o obtido usando-se o teste t.

5 — Uma experiência com um processo fabril mostrou que 2 por cento das peças produzidas por certas máquinas eram defeituosos. Suspeita-se que uma máquina esteja com defeito e por isso registra-se sua percentagem diária de peças defeituosas. Este registro de percentagem é: 2,2, 2,3, 2,1, 1,7, 3,8, 2,5, 1,9, 1,6, 1,4, 2,6, 1,5, 2,8, 2,9, 2,6, 2,5, 3,2, 4,6, 3,3, 3,0, 1,8, 2,6, 2,1, 1,8, 2,4, 2,4, 1,7, 1,6, 2,8, 3,0, 2,4. Use o teste de sinal para testar a hipótese de que a mediana da distribuição seja 2.

Seção 3

6 — Os dados seguintes representam o número de glóbulos vermelhos (milhões por milímetro cúbico) para 10 homens e 10 mulheres. (a) Usar o teste da soma de postos para ver se há alguma diferença entre os sexos com respeito aos glóbulos vermelhos. (b) Fazer o problema usando a aproximação normal para este teste.

Homens	5,02	4,58	5,57	4,52	4,84	5,36	4,27	5,15	4,93	4,72
Mulheres	4,15	4,56	3,89	4,40	4,38	4,20	4,31	4,73	4,26	3,95

7 — Explicar por que a prova dos sinais aplicada às diferenças de valores no problema 6 não seria um teste satisfatório para se aplicar a tais dados.

8 — Estes dados representam as notas em um teste de gramática para 2 grupos de estudantes. Os estudantes do grupo I graduaram-se em escolas particulares, ao passo que os estudantes do grupo II são graduados em escolas públicas. Usar o teste de soma de pontos para testar a igualdade das medianas.

I	25	30	29	34	24	25	13	32	24	30	32	37			
II	44	33	22	8	47	31	40	30	33	35	18	21	35	28	22

9 — Resolva o problema 6 por meio do teste t de Student para a diferença de 2 médias, conforme explicado no Capítulo 8, e comente os dois casos.

10 — Resolva o problema 34, Capítulo 8, por meio do teste da soma dos postos.

11 — Os dois conjuntos seguintes de valores representam os resultados da amostragem de 12 machos e 15 fêmeas após aplicação de um teste para medir o alto nível de tolerância. Use o teste da soma de postos para testar a hipótese de que os dois grupos possuam a mesma distribuição com respeito a esta medida.

M	25	30	28	34	24	25	13	32	24	30	31	35			
F	44	34	22	8	47	31	40	30	32	35	18	21	35	29	22

Seção 4

12 — Numa exposição agropecuária 2 juízes classificaram 12 potes de geléia segundo as posições seguintes:

Crítica A	6	2	9	11	7	8	10	3	12	1	4	5
Crítico B	5	7	4	10	2	11	8	6	9	3	12	1

Testar para ver se há qualquer concordância entre estes juízes por meio do coeficiente de correlação de postos.

13 — Para os dados do problema 2, substitua as notas por seus números de ordem e depois use o coeficiente de correlação de postos para testar a fim de ver se as notas de leitura foram afetadas pelo método de formar os pares de crianças.

14 — Qual deveria ser o tamanho do coeficiente de correlação de postos, para uma amostra de tamanho 20 de forma a se poder rejeitar a hipótese de não-relação?

15 — Sugira duas variáveis para as quais não seria razoável calcular o coeficiente de correlação ordinária, mas para as quais o coeficiente de correlação de ordem seria muito adequado.

13

REGRESSÃO MÚLTIPLA

1 — REGRESSÃO LINEAR MÚLTIPLA

A maioria dos problemas práticos envolve mais de uma variável para previsão. Por exemplo, se uma faculdade quiser desenvolver uma fórmula de regressão para concurso de admissão, baseada na predição do sucesso acadêmico do calouro, certamente teria de incorporar, na função de regressão, variáveis como média de notas do secundário, notas de teste de aptidão e avaliação por entrevista. Ou, na previsão de demanda por eletricidade, da parte de uma grande cidade, os planejadores teriam de considerar grande número de variáveis pertinentes, inclusive fatores como crescimento industrial, a dinâmica das futuras demandas por habitação, movimento da indústria para áreas suburbanas etc.

Os métodos para tratar com problemas de prever uma variável por meio de diversas outras variáveis, ao invés de por meio de apenas uma, são semelhantes àqueles para uma variável. Por exemplo, se alguém fosse prever a variável y em função de duas variáveis x_1 e x_2, o problema tornar-se-ia descobrir o plano que melhor se ajustasse, no sentido dos mínimos quadrados, a um diagrama de dispersão de pontos em três dimensões. A geometria de tal problema é ilustrada na Figura 1.

Como a equação de qualquer plano no espaço tridimensional determinado pelas variáveis y, x_1 e x_2 pode escrever-se na forma

$$y = a_0 + a_1x_1 + a_2x_2 \tag{1}$$

o problema é estimar os três parâmetros a_0, a_1 e a_2 pelo método dos mínimos quadrados. Neste capítulo, a notação do Capítulo 9, onde

se coloca um sinal sobre y para distinguir o valor de regressão de seu valor observado correspondente, não será usada, porque provocaria confusão, na última discussão. Deveria estar claro de seu uso se y representasse um valor observado ou um valor de regressão. Os erros de predição aqui são as distâncias verticais dos pontos ao plano. A minimização da soma dos quadrados destes erros é obtida pelas mesmas técnicas matemáticas usadas na regressão simples linear. Se há n pontos disponíveis, segue-se que os valores dos mínimos quadrados de a_0, a_1 e a_2 são obtidos resolvendo o seguinte conjunto de três equações lineares. Supõe-se que as somas sejam efetuadas sobre todos os n valores das variáveis.

$$
\begin{aligned}
a_0 n + a_1 \Sigma x_1 + a_2 \Sigma x_2 &= \Sigma y \\
a_0 \Sigma x_1 + a_1 \Sigma x_1^2 + a_2 \Sigma x_1 x_2 &= \Sigma x_1 y \\
a_0 \Sigma x_2 + a_1 \Sigma x_1 x_2 + a_2 \Sigma x_2^2 &= \Sigma x_2 y
\end{aligned} \tag{2}
$$

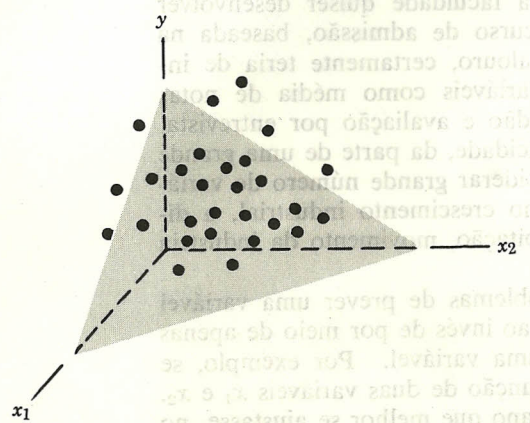


FIGURA 1. Plano de regressão em três dimensões.

Este resultado é generalizado para um número indefinido de variáveis. Assim, se alguém tivesse quatro variáveis, x_1, x_2, x_3 e x_4, por meio das quais fosse prever y, haveria cinco equações nas cinco incógnitas a_0, a_1, a_2, a_3 e a_4 para resolver.

Uma das atrações da abordagem dos mínimos quadrados nos problemas de regressão é a maneira simples de escrever as equações que devem ser resolvidas para se obter as estimativas dos parâmetros

da equação de regressão. Por exemplo, a equação (2) pode ser escrita efetuando-se as operações da maneira como segue. Primeiro, somar ambos os lados da equação (1) para obter a primeira equação dos mínimos quadrados; depois, multiplicar cada termo da equação (1) por x_1 e somar ambos os lados para obter a segunda equação dos mínimos quadrados; finalmente, multiplicar cada termo da equação (1) por x_2 e somar ambos os lados para obter a terceira equação.

Se houvesse cinco variáveis independentes ao invés de duas, na equação de regressão, teríamos de continuar este procedimento multiplicando por x_3 e somando, depois por x_4 e somando, finalmente por x_5 e somando. Haveria, então, seis equações de mínimos quadrados a serem resolvidas para se obter estimativas dos seis parâmetros incógnitos de regressão.

Com a pronta disponibilidade atual de calculadoras rápidas, é possível incorporar grande número de variáveis numa função de regressão e, portanto, aumentar nossa capacidade de prever o valor de uma variável em função de várias outras. Não é incomum usar-se 20 ou mais variáveis, quando antes os cálculos envolvidos em tais casos eram por demais pesados. Uma das evidentes vantagens de uma função de regressão que incorpora muitas das variáveis importantes relacionadas à variável y é que ela nos permite estudar como uma dessas variáveis afeta y quando todas as outras variáveis são mantidas constantes. Assim, ela permite uma análise íntima de várias relações, o que não é possível quando todas as variáveis estão mudando simultaneamente em seu quadro natural. Por exemplo, suponhamos que seja atribuído um valor fixo à variável x_2 na fórmula (1). A função de regressão afirma que, em média, o valor de y aumentará na quantidade a_1 se x_1 aumentar de uma unidade. Trata-se apenas de uma relação média que se baseia nos pressupostos semelhantes aos da regressão linear simples.

Exemplo. Como ilustração do modo como se fazem tais interpretações, consideremos o problema de regressão associado aos dados do Quadro 1. Esses dados representam os escores obtidos por 20 empregados de uma companhia em 2 testes psicológicos e seus escores de avaliação decididos previamente pela administração. O problema de regressão é determinar se o sucesso de um empregado, medido pelo seu escore de avaliação, pode ser predito com base em seus escores de testes psicológicos.

As equações dos mínimos quadrados foram calculadas como

$$
\begin{aligned}
20a_0 + 936a_1 + 837a_2 &= 1{,}364 \\
936a_0 + 50{,}380a_1 + 46{,}283a_2 &= 68{,}012 \\
837a_0 + 46{,}283a_1 + 53{,}493a_2 &= 64{,}845
\end{aligned}
$$

QUADRO 1

Notas do Teste x_1	Notas do Teste x_2	Notas de Avaliação y	Notas do Teste x_1	Notas do Teste x_2	Notas de Avaliação y
77	76	84	35	51	59
83	71	85	35	10	48
64	96	95	54	74	73
30	2	55	43	10	64
59	96	95	33	52	66
53	55	73	32	27	53
77	73	89	54	16	57
37	16	65	28	21	59
15	15	47	54	19	69
29	51	71	45	6	57

A solução destas equações é dada por $a_0 = 41{,}2$, $a_1 = 0{,}308$ e $a_2 = 0{,}302$; portanto, a equação da função da regressão linear é

$$y = 41{,}2 + 0{,}308\ x_1 + 0{,}302\ x_2 \tag{3}$$

Como se supõe que a função de regressão represente só o valor médio do y correspondente a quaisquer valores escolhidos das variáveis independentes, a interpretação dos coeficientes de regressão deve ser em termos de valores médios. Assim, se considerarmos só aqueles empregados que conseguiram, digamos, um escore de 70 no primeiro teste, a equação (3) afirma que, em média, a avaliação da administração aumentará de 0,302 para cada aumento unitário no escore do segundo teste. Por exemplo, se um empregado consegue 10 escores a mais que outro empregado no segundo teste, e tem o mesmo escore no primeiro teste, esperar-se-ia que o primeiro empregado tivesse 3 pontos a mais que o segundo na avaliação da administração. Esta interpretação não depende do valor particular atribuído a x_1. Pode-se fazer uma interpretação semelhante relativa ao coeficiente de x_2, se mantivermos fixo o valor de x_1.

Suponhamos que o valor de x_2 não fosse introduzido neste problema e que se tivesse calculado a regressão linear simples de y sobre x_1. Haveria alguma mudança notável no coeficiente de x_1 e haveria qualquer perda notável de precisão na previsão? Para responder a estas perguntas, a equação da regressão linear simples de y sobre x_1 foi calculada como sendo

$$x = 38{,}47 + 0{,}635x_1 \tag{4}$$

O termo constante a_0 mudou muito pouco; contudo, o coeficiente de x_1 quase dobrou de valor. Isto não seria muito surpreendente porque, se x_1 e x_2 tivessem o mesmo valor, a contribuição da soma dos termos x_1 e x_2 na equação (3) igualaria mais ou menos a contribuição de x_1 só na equação (4). Mas, então, surge a questão se x_1 só não poderia ser tão bom quanto x_1 e x_2 combinados na previsão de y. Para resolver esta questão, basta calcular o erro-padrão de estimativa para cada uma das funções de regressão (3) e (4). Os cálculos forneceram os valores 5,6 e 9,2. A partir desses valores fica claro que só o uso de x_1 não é tão bom quanto o uso de x_1 e x_2. Os erros de predição seriam muito menores quando usamos x_1 e x_2 do que quando usamos só x_1.

2 — TÉCNICAS DE CÁLCULO

Já que dispomos agora facilmente de calculadoras de grande velocidade, elas podem incorporar um grande número de variáveis, num problema de regressão. Mas a vantagem de uma função de regressão que contém um grande número de variáveis sobre a que contém *um pequeno número de variáveis realmente significativas* é pequena. Além disso, a dificuldade e os custos adicionais em jogo na aplicação de uma função maior podem favorecer uma função menor. O problema, portanto, é como escolher algumas variáveis importantes dentre um conjunto de variáveis, para construir uma função de regressão linear para a predição de uma dessas variáveis, y, em termos de variáveis escolhidas.

Um dos métodos de se fazer isso chama-se *regressão por passos* (*stepwise regression*) que se consegue da seguinte maneira.

Primeiro, obtém-se a equação da regressão linear de y sobre cada uma das variáveis independentes e calcula-se o erro-padrão da estimativa para cada uma dessas regressões. A variável que fornece o menor erro-padrão de estimativa é a primeira variável a ser selecionada. Depois, obtém-se para cada uma das variáveis independentes restantes a equação da regressão múltipla de y baseada em duas variáveis independentes, uma das quais deve ser a variável selecionada.

Como antes, calcula-se o erro-padrão de estimativa para cada função de regressão e seleciona-se, como a segunda variável a ser retida a variável que produz o menor de tais valores. Continua-se o processo cada vez com uma função de regressão que contenha todas as variáveis selecionadas previamente e uma das variáveis restantes, até que se tenha obtido o número de variáveis desejadas ou até que

o erro-padrão de estimativa não tenha diminuído apreciavelmente, em valor. É provável que não haja nenhuma melhoria pela incorporação de variáveis adicionais na função de regressão se não houver decréscimo apreciável no valor do erro-padrão de estimativa, em qualquer estágio.

Uma função de regressão que contenha grande número de variáveis independentes deve ser tratada com muita delicadeza, porque pode haver algumas correlações fortes entre subconjuntos dessas variáveis, e nesses casos a equação de regressão pode não ser segura. A técnica de regressão por passos provavelmente selecionará variáveis independentes altamente correlacionadas nos primeiros estágios, mas, se continuarmos, selecionará as melhores das variáveis já altamente correlacionadas. É provável que muitas variáveis que são naturais a serem selecionadas em problemas de regressão, dentro das ciências sociais, estejam correlacionadas; portanto, os pesquisadores não se devem deixar levar pelo entusiasmo provocado pelas facilidades de cálculo dos computadores. Multidões de variáveis, ao invés de ajudar, podem afastá-los da realidade e levá-los a armadilhas de interpretações injustificadas.

3 — REGRESSÃO NÃO-LINEAR

A discussão do Capítulo 9 mostrou a possibilidade de reduzir muitas relações não-lineares entre duas variáveis x e y a relações lineares, escolhendo-se as funções adequadas de x e y. Embora isso seja sempre possível teoricamente, as funções respectivas podem ser excessivamente complicadas ou talvez não seja possível determinar quais seriam as funções respectivas. Muitas vezes é, portanto, mais conveniente trabalharmos com relações não-lineares entre x e y. A curva mais simples a ser usada como modelo de regressão, se a linha reta não for suficiente, é a parábola. Sua equação pode ser escrita sob a forma

$$y = a_0 + a_1x + a_2x^2 \tag{5}$$

O problema de se ajustar uma curva deste tipo a um conjunto de pontos no plano x, y pelos mínimos quadrados é semelhante ao de ajustar uma equação de regressão múltipla a um conjunto de pontos em três dimensões. De fato, os resultados obtidos na seção 1 podem ser aplicados diretamente a este problema. Basta escolher $x_1 = x$ e $x_2 = x^2$ na equação (1) para obtermos a equação (5). Como

resultado as equações dos mínimos quadrados para a equação (5) são obtidas fazendo-se estas mesmas substituições nas equações dos mínimos quadrados dada pela equação (2). Assim, as equações cujas soluções dão as estimativas desejadas para a_0, a_1 e a_2, ajustando-se a parábola (5) a um conjunto de pontos, são as seguintes:

$$\begin{aligned} a_0n + a_1\Sigma x + a_2\Sigma x^2 &= \Sigma y \\ a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma x^3 &= \Sigma xy \\ a_0\Sigma x^2 + a_1\Sigma x^3 + a_2\Sigma x^4 &= \Sigma x^2y \end{aligned} \tag{6}$$

Com o propósito de observarmos até que ponto uma parábola se ajusta melhor a um conjunto de pontos que uma linha reta, basta calcularmos os erros de previsão para as duas curvas ajustadas e compará-lo. O erro-padrão de estimativa pode também ser calculado para ambas as curvas e comparado. Mas o erro-padrão de estimativa para uma parábola usa o divisor $n - 3$ no lugar de $n - 2$ da fórmula (6), Capítulo 9. Se houver diferença muito pequena entre os dois erros-padrão de estimativa calculados não se ganhou muito usando a parábola em vez da linha reta para a curva de regressão.

Estes métodos podem ser estendidos a curvas polinomiais de grau mais elevado. Tudo que se precisa fazer é substituir x_3 por x^3, x_4 por x^4 etc., nas formas mais gerais da equação (1) e (2) para obtermos a equação da curva de regressão polinomial de grau desejado e suas correspondentes equações de mínimo quadrado para a estimativa dos coeficientes.

Exemplo. Como ilustração do processo de ajustagem de uma parábola a um conjunto de pontos, consideremos os dados do Quadro 2 que fornecem o registro, para um período de 18 semanas, do número

QUADRO 2

x	y	x	y
242	107	200	120
255	108	210	122
261	102	237	114
268	103	284	103
275	105	293	102
282	100	298	100
222	113	302	104
214	118	306	103
230	112	313	105

de unidades x produzidas numa fábrica e do custo estimado por unidade, y, neste nível de produção.

O diagrama de dispersão para estes dados aparece na Figura 2. Parece claro que uma linha reta se ajusta mal ao conjunto de pontos, mas uma parábola ajusta-se bastante bem. Deve-se esperar aqui um tipo de relação curvilínea, porque uma fábrica que se orienta para produzir mais eficientemente, a dado nível de produção, será menos eficiente produzindo a um nível mais baixo que o ideal, e também se tentar produzir a um nível mais elevado que o planejado para sua capacidade, devido aos custos excedentes.

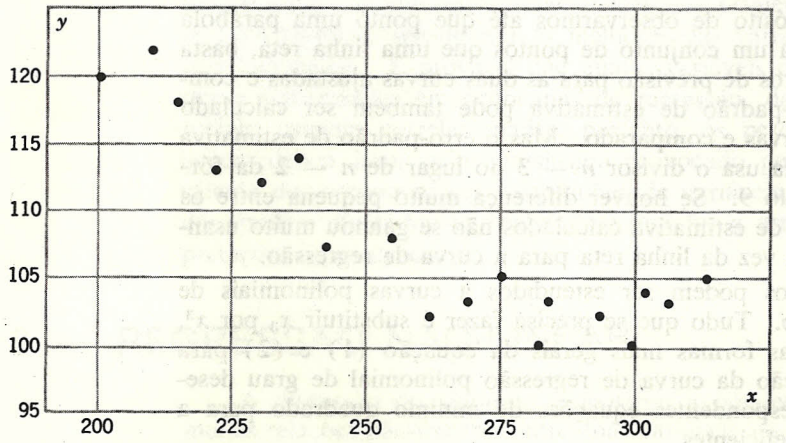


FIGURA 2. Diagrama de dispersão para o Quadro 2.

Os cálculos tornar-se-ão mais simples se, digamos, subtrairmos 200 de todos os valores de x e 100 de todos os valores de y. Façamos $X = x - 200$ e $Y = y - 100$, e escrevamos a equação da parábola

$$Y = a_0 + a_1X + a_2X^2$$

Em termos das variáveis dos dados originais isto equivale a

$$(7) \qquad y - 100 = a_0 + a_1(x - 200) + a_2(x - 200)^2$$

A parábola dos mínimos quadrados para os dados originais pode ser obtida desta equação multiplicando-se os parênteses e reunindo os termos semelhantes.

As equações (6) dos mínimos quadrados resultantes dos dados modificados são

$$\begin{aligned} 18a_0 + \quad 1092a_1 + \quad 88.250a_2 &= \quad 141 \\ 1092a_0 + \quad 88.250a_1 + \quad 7.880.538a_2 &= \quad 4.800 \\ 88.250a_0 + 7.880.538a_1 + 741.676.598a_2 &= 305.608 \end{aligned}$$

Estas são equivalente a

$$\begin{aligned} a_0 + 60{,}667a_1 + 4902{,}778a_2 &= 7{,}833 \\ a_0 + 80{,}815a_1 + 7216{,}610a_2 &= 4{,}396 \\ a_0 + 89{,}298a_1 + 8404{,}267a_2 &= 3{,}463 \end{aligned}$$

Subtraindo a primeira equação das equações restantes temos

$$\begin{aligned} 20{,}148a_1 + 2313{,}832a_2 &= -3{,}437 \\ 28{,}631a_1 + 3501{,}489a_2 &= -4{,}370 \end{aligned}$$

ou

$$\begin{aligned} a_1 + 114{,}84a_2 &= -0{,}1701 \\ a_1 + 122{,}30a_2 &= -0{,}1526 \end{aligned}$$

Resolvendo este par de equações e usando o resultado para obter a_0 de uma equação anterior achamos

$$a_2 = 0{,}002346, \; a_1 = 0{,}4395, \; a_0 = 22{,}976.$$

Substituindo estes valores na equação (7) e unindo os termos temos

$$y = 304{,}9 - 1{,}378x + 0{,}00235x^2$$

O gráfico desta equação no diagrama de dispersão da Figura 2 aparece na Figura 3. O ajustamento parece muito bom.

Para um problema onde não é tão óbvio que uma parábola dê um ajustamento melhor que uma linha reta, pode-se calcular o erro-padrão de estimativa para ambas as regressões e observar se há ganho apreciável ao usarmos uma parábola ao invés de uma linha reta. Fazendo isto, deveremos usar a fórmula (9) da próxima seção.

Como ficou indicado na seção precedente, deve-se proceder com cautela ao se interpretar uma função de regressão linear múltipla, baseada em várias variáveis independentes por causa dos possíveis efeitos de fortes correlações. Em tais problemas há a dificuldade adicional de poder o pressuposto de linearidade não ser satisfeito. Se há boa evidência, teórica ou empírica, para se crer que as médias dos

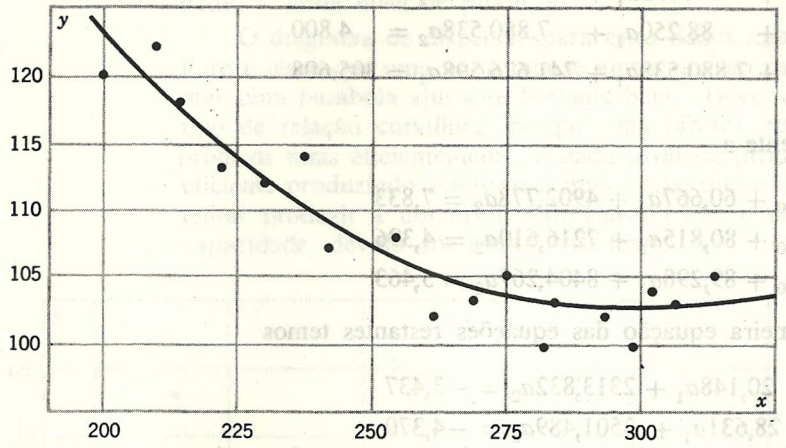


FIGURA 3. Parábola dos mínimos quadrados para o Quadro 2.

y não estão no plano, pode-se construir uma função de regressão mais geral introduzindo-se termos do segundo grau entre as variáveis independentes. Por exemplo, a função da regressão linear (1) podia ser substituída pela função de regressão não-linear

$$(8) \qquad y = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_1x_2 + a_5x_2^2$$

A técnica para se descobrir as equações dos mínimos quadrados é a mesma anterior. Basta tratar as variáveis x_1^2, x_1x_2 e x_2^2 como três novas variáveis, digamos, x_3, x_4 e x_5 e escreveremos as seis equações dos mínimos quadrados correspondentes. Como a inclusão de termos do segundo grau aumenta rapidamente o número de termos numa função de regresso, raramente se usa esta técnica a não ser que o número de variáveis independentes seja muito pequeno.

Como foi assinalado no Capítulo 9, muitas vezes é possível reduzir uma relação não-linear entre duas variáveis a uma relação linear aproximada ou exata, escolhendo-se as funções adequadas das duas variáveis como as variáveis em estudo. Às vezes, também, é possível fazer isso em problemas de regressão múltipla. Pode-se empregar esta mesma técnica para simplificar as relações que, de outro modo, poderiam ser muito complicadas. O problema seguinte é um exemplo de como uma função logarítmica operou tal simplificação.

Exemplo. Os recordes mundiais de corrida de 1973 ou 1974, para distâncias de 100 jardas a 10 milhas aparecem no Quadro 3. Este quadro lista também a velocidade do corredor para a distância correspondente em jardas por segundo.

QUADRO 3

Distância	Tempo	Velocidade
100 jardas	9,0	11,1
220 jardas	19,5	11,3
440 jardas	44,5	9,9
880 jardas	1:44,6	8,4
1 milha	3:51,1	7,6
2 milhas	8:13,8	7,1
3 milhas	12:47,6	6,9
6 milhas	26:47,0	6,6
10 milhas	46:37,8	6,3

A velocidade média em 100 jardas é menor que em 200 jardas, porque uma grande fração do tempo da corrida é consumida na recuperação da velocidade a partir de um ponto de partida. Por causa deste fato omitiu-se, na análise que segue, os dados relativos às 100 jardas.

Para observarmos a relação entre distância e velocidade, é útil calcularmos o logaritmo da distância e marcarmos a velocidade contra o logaritmo da distância. O Quadro 4 fornece a relação entre velocidade e logaritmo da distância, na base 10.

QUADRO 4

Velocidade	11,3	9,9	8,4	7,6	7,1	6,9	6,6	6,3
Log da distância	2,34	2,64	2,94	3,25	3,55	3,72	4,02	4,25

Consideremos agora as soluções dos seguintes problemas.

(a) Trace o gráfico dos pontos correspondentes à relação.

(b) Ache a equação dos mínimos quadrados (polinomial do 3.º grau) que se ajustará ao conjunto de pontos obtidos em (a), e esboce-a sobre o gráfico de (a).

(c) Supondo que a função de regressão em (b) seja também válida para a "maratona" (aproximadamente 26,2 milhas), estime a velocidade para a maratona.

(d) Com base na resposta (c), estime o recorde mundial para a maratona. O recorde atual é de 2 horas, 12 minutos e 11,2 segundos.

(a) O gráfico aparece na Figura 4.

(b) Os cálculos baseados num conjunto de quatro equações com 4 incógnitas, semelhantes aos da equação (6) dão a equação

$$y = 49{,}9921 - 29{,}6962x + 6{,}9433x^2 - 0{,}5583x^3$$

O gráfico desta equação aparece na Figura 4.

(c) 26,2 milhas = 26,2 (1760) jardas = 46,112 jardas. Log 46,112 = 4,6638 = x para a maratona. Para este valor de x, o valor correspondente de y, dado pela fórmula em (b), é

y (maratona) = 5,8833 = velocidade em jardas por segundo.

(d) $$\text{Tempo} = \frac{\text{distância}}{\text{velocidade}} = \frac{46.112}{5{,}8833} = 7837{,}8 \text{ segundos} =$$

= 2 horas, 10 minutos, 38 segundos;

o que está muito próximo do recorde real, com um erro de cerca de 1 por cento. Uma parábola aqui teria sido insatisfatória, porque começaria aumentando para valores de x além do ponto mínimo, implicando velocidade crescente para longas distâncias.

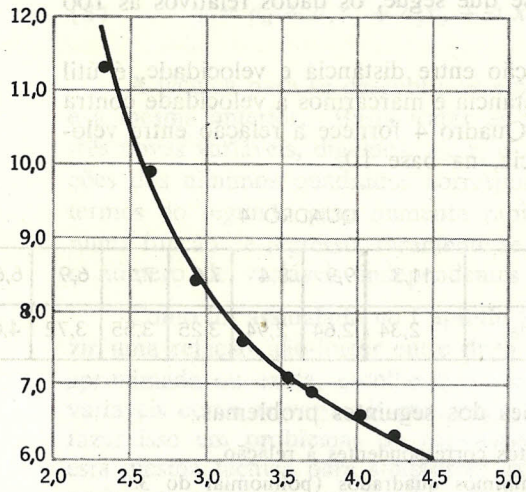


FIGURA 4. Gráfico da velocidade contra o logaritmo da distância.

4 — CONFIANÇA

As fórmulas para determinar o grau de confiança de uma função de regressão múltipla ou não-linear são semelhantes às usadas para o caso de regressão linear simples. O erro-padrão de estimativa é um pouco modificado para compensar o número de coeficientes que precisam ser estimados. É definido assim:

(9) $$s_e = \sqrt{\frac{\sum_{j=1}^{n}(y_j - y'_j)^2}{n-k}}$$

onde k é o número de coeficientes, inclusive o termo constante, na fórmula da regressão e onde y'_j expressa o valor da função de regressão correspondente a y_j. Por exemplo, k seria 3 para as fórmulas (1) e (5) e seria 6 para a fórmula (8). Sob a mesma espécie dos pressupostos da regressão linear simples, o erro-padrão de estimativa pode ser usado para indicar que tamanho de erros podem ser esperados ao se predizer um valor de y por meio de uma função de regressão.

As fórmulas para se determinar a confiabilidade dos coeficientes de regressão como estimativas dos coeficientes teóricos correspondentes são muito mais complexas que as da regressão simples. Alguns dos programas de computação para regressão múltipla imprimem informações que podem ser usadas para resolver estes problemas de confiabilidade. Em particular, eles nos habilitam a testar se algum coeficiente especificado tem um valor zero e, dessa forma, se a variável correspondente é de pouco uso numa fórmula de regressão. Naturalmente imprimem também a equação da função dos mínimos quadrados, quer se trate de um problema de regressão múltipla quer de um problema de ajustamento de curva.

Mas tudo isso requer acesso a um computador digital capaz de manipular tais programas.

Em resumo, as técnicas empregadas na regressão linear simples podem também ser usadas na regressão linear múltipla ou na regressão não-linear, mas com muito mais cautela, porque os pressupostos que as justificam são menos prováveis de ocorrer em problemas reais.

EXERCÍCIOS

1 — Diante dos dados abaixo use a fórmula (2) para encontrar a equação da função de regressão linear de **y** sobre $\mathbf{x_1}$ e $\mathbf{x_2}$.

x_1	0	1	2	3	4	5	6	7	8	9
x_2	2	10	6	12	4	11	15	14	9	13
y	0	4	9	18	10	16	20	20	18	25

2 — Diante dos seguintes dados, ache a equação da função de regressão linear de **y** sobre $\mathbf{x_1}$ e $\mathbf{x_2}$

x_1	2	1	5	8	7	2	1	3	0	9	4	6
x_2	6	5	5	7	3	1	8	2	6	7	7	9
y	13	9	15	16	21	9	15	11	12	30	19	22

3 — Para os dados do problema 1, (a) ache a equação da função da regressão linear de **y** sobre $\mathbf{x_1}$; (b) calcule a soma dos quadrados dos erros de estimativa e compare-a com a soma correspondente baseada na função encontrada no problema 1 para ver se a variável $\mathbf{x_2}$ foi beneficiada.

Seção 3

4 — Os seguintes dados dão o número de homens/hora (**y**) exigidos para se completar uma tarefa como função do número de unidades já completadas. (a) Faça o gráfico dos dados. (b) Ajuste uma parábola ao conjunto de pontos em (a) por meio da fórmula (6).

x	0	1	2	3	4	5
y	40	34	30	24	22	20

Explique por que uma parábola pode ser de um tipo não-realista de curva para usar como modelo num problema deste tipo.

5 — Diante dos seguintes dados, (a) dispor os pontos; (b) encontrar a equação da reta dos mínimos quadrados; (c) encontrar a equação da parábola dos mínimos quadrados. Será que o ajustamento pela parábola é muito melhor?

x	5	6	7	8	9	10	11	12	13	14	15	16	17
y	7.6	9.5	9.3	10.3	11.1	12.1	13.3	12.7	13.0	13.8	14.6	14.6	14.7

Seção 4

6 — Use a fórmula (9) e os resultados dos cálculos do problema 3 para achar o erro-padrão de estimativa para ambas as funções de regressão desse problema. Que conclusões tirar?

7 — Diante dos seguintes dados, (a) ache a equação do plano dos mínimos quadrados; (b) calcule os erros de previsão; (c) calcule o desvio-padrão da estimativa; (d) omita a variável $\mathbf{x_2}$ e depois ache a equação da reta dos mínimos quadrados; (e) calcule o desvio-padrão de estimativa para a linha obtida em (d) e compare com o valor obtido em (c). Ganhou-se muito com a introdução da variável $\mathbf{x_2}$ além da $\mathbf{x_1}$?

x_1	1	7	6	2	5	2	8	3	7	6
x_2	4	6	8	0	1	5	8	8	3	0
y	11	21	23	7	13	18	30	18	21	20

8 — Calcule o erro-padrão de estimativa para ambas as funções de regressão do problema 5. A parábola se ajusta muito melhor que a linha reta?

Geral

Os dados da Tabela XII do Apêndice podem ser usados para construir vários problemas de regressão múltipla. Por exemplo, poder-se-ia tentar prever a nota de avaliação de um professor ($\mathbf{x_5}$) com base na nota total do estudante no curso ($\mathbf{x_3}$) e na nota média do mesmo ($\mathbf{x_4}$). Muitos dos dados podem ser usados como desejados, dependendo das facilidades de cálculo disponíveis.

14

DECISÃO ESTATÍSTICA

1 — INTRODUÇÃO

Nos capítulos precedentes enfatizou-se a inferência estatística e deu-se pouca atenção à interpretação e tomada de decisão. Isto é devido ao fato de a decisão a ser tomada após a inferência depender de considerações exógenas, não disponíveis ao estatístico.

O empresário que anima uma pesquisa, por exemplo, deve usar sua capacidade de julgamento ao decidir como usar melhor os resultados desta pesquisa.

Mas, há certos tipos de problemas nos quais o estatístico pode indicar qual a melhor decisão a ser tomada, desde que se aceitem as premissas nas quais se baseiam. O mais simples de tais problemas é aquele onde só há dois estados ou eventos possíveis e duas ações. Por exemplo, um agricultor pode estar preocupado com a possibilidade de uma forte geada arruinar sua plantação, e deve decidir se a protege ou não contra esta eventualidade. Os dois estados ou eventos possíveis, aqui, são a ocorrência ou não-ocorrência de geada, e as duas ações possíveis são proteger-se ou não contra ela. Uma versão mais geral deste tipo de problema ocorre quando há, digamos, $k \geq 2$ eventos possíveis e $l \geq 2$ ações possíveis. Por exemplo, uma empresa pode classificar um novo produto em uma das três categorias: produto de grande sucesso, moderado, ou produto fracassado. Sua escolha de ações pode consistir em propaganda em televisão, propaganda em rádio, propaganda em jornal ou propaganda em folhetos. Como os métodos que estão para ser propostos não dependem dos valores de k e l, desde que excedam um, bastará estudar a versão mais simples do problema no qual $k = 2$ e $l = 2$.

2 — TABELA "PAYOFF"

Para explicarmos o raciocínio que se esconde atrás da solução proposta dos problemas do tipo precedente, consideremos primeiro o seguinte problema particular, semelhante ao discutido na seção 4 do Capítulo 4. Suponhamos que o promotor de um evento esportivo esteja duvidando se faz ou não seguro contra chuva. Vamos supor que seu lucro líquido seja de \$ 20.000 se não chover, mas apenas de \$ 2.000 se chover e que ele esteja estudando uma apólice de seguro de \$ 20.000 ao custo de \$ 5.000. Esta informação aparece bem na tabela seguinte chamada tabela *payoff* que lista seus lucros líquidos sob várias possibilidades (Quadro 1).

QUADRO 1

Tempo	Ação	
	Seguro	Nenhum seguro
Claro	15.000	20.000
Chuvoso	17.000	2.000

A base da preferência de uma dessas ações ou decisões sobre as outras será o lucro esperado. Assim, a decisão de fazer seguro será tomada se, e somente se, o lucro esperado ao fazê-lo for maior que o lucro esperado em não fazê-lo. Para calcularmos estes dois lucros esperados façamos p simbolizar a probabilidade de que o tempo esteja claro no dia do evento esportivo. Então $1 - p$ será a probabilidade de que choveria nesse dia. Recordando a definição de expectativa dada no Capítulo 4, os cálculos dos lucros esperados sob as duas ações possíveis, denotadas por E_I e E_N, podem ser efetuados como mostra o Quadro 2.

QUADRO 2

Tempo	Probabilidade	Seguro		Nenhum Seguro	
		Lucro	Lucro Esperado	Lucro	Lucro Esperado
Claro	p	15.000	15.000p	20.000	20.000p
Chuvoso	1 — p	17.000	17.000(1 — p)	2.000	2.000(1 — p)
Totais	1		E_I		E_N

Dessa forma

(1) $$E_I = 15.000p + 17.000(1 - p)$$

e

(2) $$E_N = 20.000p + 2000(1 - p)$$

Suponhamos que o valor de p seja desconhecido. Então estas duas ações possíveis serão igualmente boas desde que p satisfaça a equação obtida fazendo $E_I = E_N$. Assim, p deve satisfazer a equação

$$15.000p + 17.000(1 - p) = 20.000p + 2000(1 - p)$$

Vê-se facilmente que a solução desta equação é $p = 0{,}75$. O valor de p que produz igualdade entre os dois valores esperados de um problema como este chama-se *ponto de equilíbrio.*

É claro intuitivamente que se $p > 0{,}75$ o promotor não faria seguro, porque quanto melhor o tempo tanto menor a necessidade de seguro, e as duas ações possíveis são igualmente boas quando $p = 0{,}75$. Este fato é prontamente verificado algebricamente, resolvendo a inequação $E_N > E_I$, isto é,

$$20.000p + 2000(1 - p) > 15.000p + 17.000(1 - p)$$

Como as inequações são manipuladas da mesma maneira que as igualdades, exceto quando se manipula por um número negativo, esta inequação se reduz a

ou $$20p + 2(1 - p) > 15p + 17(1 - p)$$

$$5p > 15(1 - p)$$

ou $$p > \frac{15}{20}$$

Em vista dos cálculos precedentes, o promotor de esportes deveria consultar os registros meteorológicos para o período de tempo em que deve realizar-se o evento para ver se a proporção de dias claros (não chuvosos) excede 0,75. Se exceder e se ele estiver querendo usar o lucro esperado para a tomada de decisões, não deve fazer seguro. Isto supõe que o evento deve realizar-se num futuro tão longínquo que não se dispõe de nenhuma previsão de tempo garantida e que as taxas de seguro se baseiam também neste fato. Geralmente as taxas de seguro se baseiam em experiência passada, mas poderiam basear-se numa previsão se o evento devesse ocorrer dentro de um intervalo de tempo suficientemente curto para que uma previsão seja possível.

Suponhamos que os registros meteorológicos para este período de tempo tenham revelado que 80 por cento dos dias serão claros. A decisão correta neste caso é não fazer seguro. O problema então é saber qual o lucro excedente esperado pelo promotor se não fizer seguro (isto é, quanto vai ganhar a mais do que se tivesse feito seguro). Cálculos onde entram em jogo as fórmulas (1) e (2) dão os valores

$$E_I = 15.400 \quad \text{e} \quad E_N = 16.400$$

Dessa forma, o promotor pode esperar que seus lucros sejam $ 1.000 mais elevados se escolher a decisão correta. Para um promotor que sente não poder enfrentar o risco, ganhando só $ 2.000 se chover, a filosofia de basear as decisões no lucro esperado pode não ser atraente aqui.

3 — AVALIAÇÃO DA INFORMAÇÃO

Em vista da discussão precedente, surge a questão de como utilizar a informação vinda da experiência ou do julgamento para melhorar a precisão da tomada de decisão. Para o problema da seção 2 não há muito a ser feito para melhorar a estimativa de p obtida pela consulta dos registros, a não ser que se espere até um pouco antes do evento e se obtenha uma previsão segura. Se as taxas de seguro não mudarem, então, tal informação poderia ser valiosa. Para problemas que envolvem probabilidades que podem ser estimadas a partir de amostras, o processo de tomada de decisão pode ser melhorado utilizando-se informações de tais amostras.

Esta seção considerará a avaliação de informação não-amostral e a próxima estudará a avaliação de informação obtida de amostras. No problema da seção 2, a probabilidade de tempo bom baseava-se nos registros meteorológicos passados para a data anual e forneciam o valor $p = 0{,}8$. Suponhamos que um meteorologista se proponha a fornecer ao promotor uma previsão por $ 200, uma semana antes do acontecimento. Os registros do meteorologista revelam que ele está correto em tais previsões 75 por cento das vezes. Portanto, o problema é determinar se o custo da previsão é justificado pelo aumento dos lucros. Há agora três ações possíveis, em vez de duas, porque o promotor pode comprar o aviso do meteorologista e fazer o seguro quando a previsão for de chuva. As duas possibilidades

são como antes, fazer ou não fazer seguro, sem comprar a previsão. Em vez de termos de tratar com duas situações possíveis, temos de considerar agora quatro situações de previsão de tempo. Estas aparecem na tabela *payoff* do Quadro 3.

QUADRO 3

Evento	Probabilidade	Ação		
		Seguro	Nenhum Seguro	Lucro Esperado
Claro, previsão de claro	(0,8) (0,75) = 0,60	15.000	20.000	20.000
Claro, previsão de chuva	(0,8) (0,25) = 0,20	15.000	20.000	15.000
Chuvoso, previsão de claro	(0,2) (0,25) = 0,05	17.000	2.000	2.000
Chuvoso, previsão de chuvoso	(0,2) (0,75) = 0,15	17.000	2.000	17.000

O cálculo do lucro esperado sob essas três ações possíveis é efetuado da maneira usual e origina os três valores esperados seguintes.

$$E_I = 15.000\ (0,60) + 15.000\ (0,20) + \\ + 17.000\ (0,05) + 17.000\ (0,15) = 15.400$$

$$E_N = 20.000\ (0,60) + 20.000\ (0,20) + \\ + 2.000\ (0,05) + 2.000\ (0,15) = 16.400$$

$$E_F = 20.000\ (0,60) + 15.000\ (0,20) + \\ + 2.000\ (0,05) + 17.000\ (0,15) = 17.650$$

Os valores de E_I e E_N são, naturalmente, os mesmos que os valores correspondentes obtidos na seção 2 para $p = 0,8$, porque as previsões do meteorologista não foram usadas para essas duas ações possíveis. Como o melhor antes era não fazer seguro, o ganho no lucro esperado pela utilização dos serviços do meteorologista é dado por

$$17.650 - 16.400 = 1250$$

Destes cálculos segue-se que o aviso do meteorologista vale \$ 1.250 para o promotor, em termos de lucro esperado e, portanto, ele deveria estar feliz de só pagar \$ 200 pelo aviso. Supôs-se nestes cálculos que a seguradora quis vender seguros uma semana antes do acontecimento, conforme as taxas especificadas, e, portanto, não comprou os serviços do meteorologista para determinar novas taxas.

O procedimento que acabamos de descrever é aplicável a muitas situações nas quais o problema é determinar se os lucros esperados podem ser aumentados pela compra de informações e se o ganho nesses lucros esperados excede o custo da obtenção da informação.

4 — UTILIZAÇÃO DE NOVA INFORMAÇÃO: AMOSTRAGEM

A seção precedente considerou o problema de como utilizar e avaliar a informação obtida de fontes que não fossem a amostragem. Nesta seção o problema é determinar como melhorar um processo de tomada de decisão por meio da amostragem.

Com a finalidade de mostrarmos como se procede para conseguirmos isto, apresentaremos o seguinte problema artificial. Após sua solução, aplicaremos os mesmos métodos na resolução de um problema empresarial real.

Exemplo 1. Uma caixa contém três moedas, duas delas são regulares e a terceira tem cara de ambos os lados. Devemos retirar uma única moeda da caixa. Trata-se de decidir se a moeda retirada é regular ou irregular (duas caras) se a tabela *payoff* para as possibilidades é a do Quadro 4.

QUADRO 4

Tipo de Moeda	Ação	
	Decisão pela Regular	Decisão pela de Duas Caras
Regular	9	−6
Com duas caras	−12	15

Os cálculos necessários na determinação dos lucros esperados sob as duas decisões possíveis são efetuados de maneira usual e mostrados no Quadro 5.

QUADRO 5

Tipo de Moeda	Probabilidades Anteriores	Decisão pela regular		Decisão pelas duas caras	
		Lucro	Lucro Esperado	Lucro	Lucro Esperado
Regular	$\frac{2}{3}$	9	6	−6	−4
Com duas caras	$\frac{1}{3}$	−12	−4	15	5
Totais	1		2		1

Como o lucro esperado é 2 se decidirmos em favor da moeda regular, e somente 1 se decidirmos em favor da moeda de duas caras, deveríamos decidir retirar a moeda regular.

Consideremos agora como se procede quando se obtém uma informação adicional a partir de um experimento amostral. Em especial, suponhamos que a moeda retirada seja lançada três vezes e se obtenham três caras.

Esta informação adicional pode ser incorporada à análise calculando-se um novo conjunto de probabilidades para substituir as probabilidades da segunda coluna do Quadro 5. As últimas probabilidades, a saber 2/3 e 1/3 chamam-se *probabilidades anteriores* para este problema, porque são as probabilidades disponíveis antes da realização do experimento amostral. Após a obtenção dos resultados do experimento amostral, é possível usar a fórmula de Bayes do Capítulo 3 para se conseguir estimativas melhoradas para estas probabilidades. Tais probabilidades são chamadas *probabilidades posteriores,* porque são probabilidades que se aplicam depois da realização do experimento e do uso dos resultados amostrais no cálculo de novas probabilidades. Se os dois eventos da primeira coluna do Quadro 5 forem simbolizados por R e T, as probabilidades anteriores serão dadas por

$$P\{R\} = \frac{2}{3} \quad \text{e} \quad P\{T\} = \frac{1}{3}$$

Seja S o resultado do experimento amostral, isto é, o resultado de se obter três caras nos três lançamentos da moeda. Neste caso as probabilidades posteriores serão dadas pelos símbolos

$$P\{R|S\} \quad \text{e} \quad P\{T|S\}$$

Estas probabilidades posteriores podem ser calculadas usando-se a fórmula de Bayes sob a modalidade

$$P\{R|S\} = \frac{P\{R\}P\{S|R\}}{P\{R\}P\{S|R\} + P\{T\}P\{S|T\}}$$

Embora o valor de $P\{T|S\}$ possa ser obtido subtraindo-se $P\{R|S\}$ de 1, é aconselhável calcular $P\{T|S\}$ diretamente da fórmula correspondente de Bayes para nos guardarmos contra possíveis erros de cálculo.

As probabilidades condicionais, $P\{S|R\}$ e $P\{S|T\}$, necessárias aqui, estão baseadas sobre três sucessos em três tentativas independentes e são dadas por

$$P\{S|R\} = \frac{1}{8} \quad \text{e} \quad P\{S|T\} = 1$$

Conseqüentemente

$$P\{R|S\} = \frac{\frac{2}{3}\cdot\frac{1}{8}}{\frac{2}{3}\cdot\frac{1}{8} + \frac{1}{3}\cdot 1} = \frac{1}{5}$$

e

$$P\{T|S\} = \frac{\frac{1}{3}\cdot 1}{\frac{2}{3}\cdot\frac{1}{8} + \frac{1}{3}\cdot 1} = \frac{4}{5}$$

Agora a análise prossegue como antes com as probabilidades anteriores da segunda coluna do Quadro 5 substituídas pelas probabilidades posteriores já calculadas.

Estes cálculos aparecem no Quadro 6.

Agora é necessário inverter a decisão anterior favorável à moeda regular na decisão segundo a qual a moeda retirada tem duas caras. A diferença de expectativa é agora bastante pronunciada quando

QUADRO 6

Tipo de Moeda	Probabilidades Posteriores	Decisão pela Regular		Decisão pelas duas caras	
		Lucro	Lucro Esperado	Lucro	Lucro Esperado
Regular	$\frac{1}{5}$	9	1,8	−6	−1,2
Com duas caras	$\frac{4}{5}$	−12	−9,6	15	12
Totais	1		−7,8		10,8

comparada à diferença anterior antes de se concluir o experimento amostral.

Exemplo 2. Os métodos empregados para resolver o problema artificial anterior serão agora usados para resolver um problema real. O fabricante de um produto bastante complexo, como um automóvel, defronta-se com o problema de como testar e inspecionar o produto na fábrica antes de expedi-lo ao intermediário. Ele pode fazer uma inspeção superficial e concordar em pagar ao intermediário os custos de reparos provenientes das reclamações dos clientes ou sujeitar o produto a testes e inspeções, na própria fábrica, para eliminar a maioria das reclamações que seriam feitas aos intermediários.

Suponhamos que o produto seja um automóvel e que seus custos de teste e inspeção para verificação de peças e montagem defeituosa suba a $ 24, com um adicional de $ 20, em média, para conserto do carro defeituoso. Suponhamos além disso que custe, em média, $ 80 por carro, ao intermediário, a correção de defeitos encontrados por clientes quando a fábrica não efetua cuidadosamente o programa anterior de teste-inspeção. Vamos supor também que os inspetores de fábrica e os clientes detectem a mesma proporção de carros defeituosos.

Vamos supor agora que o fabricante tenha conservado registros semanais da porcentagem de carros defeituosos surgidos na linha de produção. Expressemos esta experiência na forma do Quadro 7.

QUADRO 7

P_i	0,2	0,3	0,4	0,5	0,6
$P\{p_i\}$	0,2	0,2	0,3	0,2	0,1

Um valor particular de $P\{p_i\}$ é a proporção dessas semanas quando a fração de carro defeituosos for o valor p_i correspondente. Por exemplo, em 10 por cento desses períodos semanais descobriu-se que 60 por cento dos carros tinham um ou mais defeitos. Tais frações e proporções estão expressas em forma decimal, e arredondadas.

Se se supõe que a qualidade da produção corrente é semelhante à baseada em experiência passada, os valores do Quadro 7 podem ser usados para o cálculo das perdas esperadas sob as duas ações possíveis da inspeção da fábrica (F) ou da queixa do cliente (C). Mas ao fazermos isto entende-se que não se está considerando fatores importantes como espaço de fabricação, requisitos de programação e relações cliente-intermediário. Estes são fatores intangíveis que surgem em qualquer processo de tomada de decisão, mas muito complexos para serem considerados aqui.

O cálculo das perdas esperadas consiste em avaliar tais perdas para cada valor de p_i e somá-las. Supondo que se produza N carros num fluxo de produção semanal e que uma fração p_i deles contenha um ou mais defeitos, a perda do fabricante devida à inspeção será de $24N$ unidades monetárias e para o reparo dos Np_i carros defeituosos será de $20Np_i$; dessa forma, a perda total será

$$L_F = 24N + 20NP_i$$

Se não se fizesse nenhuma inspeção, estes carros defeituosos custariam ao fabricante $80Np_i$ unidades monetárias, pagas ao intermediário; dessa forma.

$$L_c = 80Np_i$$

O cálculo destas perdas esperadas para os vários valores de p_i aparecem no Quadro 8.

Com base nesses cálculos, a decisão mais correta, no caso, é deixar que o intermediário corrija os defeitos e reembolse-o depois pelos custos.

Se houver razão para crer que a qualidade da produção corrente difere um pouco dos fluxos de produção passados, o processo ante-

QUADRO 8

p_i	$P\{p_i\}$	Inspeção da Fábrica		Reclamação do Cliente	
		Perda	Perda Esperada	Perda	Perda Esperada
0,2	0,2	28N	5,6N	16N	3,2N
0,3	0,2	30N	6,0N	24N	4,8N
0,4	0,3	32N	9,6N	32N	9,6N
0,5	0,2	34N	6,8N	40N	8,0N
0,6	0,1	36N	3,6N	48N	4,8N
Totais	1		31,6N		30,4N

rior poderá ser melhorado tomando-se uma amostra da produção corrente e incorporando-a ao processo de tomada de decisão por meio da fórmula de Bayes, da mesma maneira que o problema artificial da moeda. Para ilustrarmos como se faz isto, suponhamos que se tome da produção normal uma amostra aleatória de 50 carros e que, após testes e inspeção cuidadosa na fábrica, encontrem-se 25 defeituosos sob algum aspecto. Estes 50 carros podem ser tratados como 50 tentativas de um experimento para o qual a probabilidade de sucesso (aqui a descoberta de um carro defeituoso) é p e para o qual ocorreram 25 sucessos. Como resultado, pode-se aplicar aqui a distribuição binomial a este problema amostral de modo que ela presida ao cálculo das probabilidades posteriores. As probabilidades, $P\{p_i\}$, encontradas na segunda coluna do Quadro 8, são as probabilidades anteriores para este problema. Se S significa o resultado da obtenção de 25 sucessos em 50 tentativas do experimento amostral, as probabilidades posteriores desejadas necessárias para substituir as probabilidades anteriores são as probabilidades $P\{p_i \mid S\}$. A versão da fórmula de Bayes para este problema é portanto

$$(3) \qquad P\{p_i \mid S\} = \frac{P\{p_i\}P\{S \mid p_i\}}{\sum_{i=1}^{5} P\{p_i\}P\{S \mid p_i\}}$$

Visto que a distribuição binomial é aplicável a este experimento amostral, as probabilidades condicionais necessárias a esta fórmula são dadas por

$$(4) \qquad P\{S \mid p_i\} = \frac{50!}{25!\,25!} p_i^{25}(1 - p_i)^{25}$$

Esta fórmula, quando substituída na fórmula (3) para cada valor p_i no Quadro 8, produzirá as probabilidades posteriores desejadas necessárias para substituir as probabilidades anteriores, $P\{p_i\}$, do Quadro 8. Como ilustração, consideremos o cálculo de $P\{p_3 \mid S\}$. Da fórmula (3) e (4) segue-se que

$$P\{p_3 \mid S\} = \frac{(0,3)\,\frac{50!}{25!\,25!}\,(0,4)^{25}\,(0,6)^{25}}{\sum_{i=1}^{5} P\{p_i\}\,\frac{50!}{25!\,25!}\,(p_i)^{25}\,(1 - p_i)^{25}}$$

$$= \frac{0,3\,(0,4)^{25}\,(0,6)^{25}}{\sum_{i=1}^{5} P\{p_i\}\,p_i^{25}\,(1 - p_i)^{25}}$$

$$= \frac{(0,3)\,(0,24)^{25}}{\sum_{i=1}^{5} P\{p_i\}\,[p_i\,(1 - p_i)]^{25}} = 0,312$$

Tais quantidades podem ser facilmente obtidas por calculadoras eletrônicas manuais. Os resultados de tais cálculos podem ser vistos na segunda coluna do Quadro 9. Este quadro mostra também os cálculos necessários para se obter as perdas esperadas baseadas nas probabilidades posteriores.

QUADRO 9

P_i	$P\{p_i \mid S\}$	Inspeção da Fábrica		Reclamação do Cliente	
		Perda	Perda Esperada	Perda	Perda Esperada
0,2	0,000	28N	0,00N	16N	0,00N
0,3	0,007	30N	0,21N	24N	0,17N
0,4	0,312	32N	9,98N	32N	9,98N
0,5	0,577	34N	19,62N	40N	23,08N
0,6	0,104	36N	3,74N	48N	4,99N
Totais	1		35,55N		38,22N

Observe que a informação fornecida pela amostra inverteu a decisão antes recomendada. Por causa da alta proporção de carros defeituosos na amostra, parece agora mais econômico efetuar uma inspeção cuidadosa na fábrica, ao passo que, antes de se tomar a amostra, o inverso era verdadeiro. A qualidade de produção parece ter-se deteriorado.

EXERCÍCIOS

Seção 2

1 — Vamos supor que o leitor obtenha a licença de vender sorvete ou "cachorro-quente" num jogo de **baseball**. A quantidade de lucro a ser obtido por cada uma dessas licenças em função da meteorologia aparece na tabela **payoff**, que segue. Se as probabilidades associadas aos três tipos de tempo forem 0,6, 0,3 e 0,1, respectivamente, que licença escolheria?

	Licença	
Tempo	Sorvete	Cachorro-quente
Ensolarado	300	200
Nublado	100	200
Chuvoso	—50	100

2 — Apague a primeira linha do problema 1 e ache depois uma probabilidade para tempo claro de tal modo que as duas licenças sejam igualmente boas, supondo que não chova.

3 — Um homem de negócios tenta decidir se investe nos casos **A** ou **B**. Ele julga que as probabilidades de sucesso nos dois casos seja de 0,2 e 0,4 respectivamente. Se o investimento **A** for bem sucedido, seu lucro será de $ 40.000. Se falhar, perderá $ 10.000. Se o caso **B** for bem sucedido, seu lucro será de $ 20.000, mas no caso de fracasso a perda será de $ 12.000. Que caso deveria escolher?

4 — Um agricultor deve decidir quando plantar uma colheita sensível à geada em seu primeiro estágio de crescimento. Se plantar em 1 de março e não ocorrer nenhuma geada depois, poderá vender sua primeira colheita em amadurecimento conseguindo um lucro de $ 10.000. Se houver geada, perderá $ 2.000 de trabalho e semente. Esperando duas semanas antes de plantar, a colheita posterior fornecerá um lucro de $ 5.000. Se a probabilidade de geada após 1 de março for de 0,6 e após 15 de março de 0,2 e o agricultor desejar maximizar o lucro esperado, que data deveria escolher?

5 — Um jornaleiro está indeciso sobre quantos jornais ordenar. O quadro seguinte dá as probabilidades (estimadas) na venda do número correspondente de jornais. Suponha por simplicidade que as ordens devem vir em unidades de 100 e que as vendas ocorram sempre em unidades de 100. Vamos supor também que pelo menos 100 jornais serão sempre vendidos mas nunca mais de 300. Se cada jornal lhe custar 6 unidades monetárias e ele o vender por 10 e se os jornais não vendidos não tiverem preço, determine quantos jornais ele deveria ordenar para maximizar o lucro esperado.

Número	100	200	300
Probabilidade	0,3	0,5	0,2

6 — O comprador de uma peça cara de equipamento eletrônico está indeciso sobre quantos tubos sobressalentes deve comprar, ao adquirir o equipamento. Cada tubo sobressalente lhe custará $ 50 por ocasião da compra da unidade, mas lhe custará $ 150 incluindo custos de mora se comprado mais tarde quando for preciso fazer a reposição. Os tubos não usados são considerados sem valor. Os registros do fabricante mostram a seguinte experiência sobre as necessidades dos compradores deste tipo de equipamento.

(a) Construa uma tabela de custos **payoff** 4 por 4 sobre as 4 possíveis necessidades e sobre as 4 compras possíveis.

(b) Tratando o quadro de freqüências relativas como probabilidades, determine quantos tubos sobressalentes devem ser comprados para minimizar o custo esperado das substituições.

Número de substituições	0	1	2	3
Freqüência relativa	0,4	0,3	0,2	0,1

7 — Uma distribuidora de discos está vendo se patrocina o disco de um novo grupo vocal. Fará 50 000 discos e o distribuirá pelo país. O custo de fabricação e distribuição de cada disco é de $ 2. Esses discos são vendidos ao preço unitário de $ 4. A experiência com casos semelhantes mostra que em 30 por cento de tais patrocínios vende-se 40 por cento dos discos, em 30 por cento deles 50 por cento dos discos são vendidos, e em 40 por cento deles vende-se 60 por cento de tais discos. Supondo que estas sejam todas as possibilidades possíveis determine se a companhia deve empresar este disco.

Seção 3

8 — Certo homem de negócios está vendo se investe $ 10.000 numa produção de petróleo. A experiência passada indica que só em 20 por cento de tais casos ter-se-á óleo suficiente para gerar lucro. A tabela **payoff** abaixo fornece os lucros médios esperados para um investimento de $ 10.000.

	Ação	
Tipo de Poço	Investir	Não Investir
Produtivo	15.000	0
Não-produtivo	—5.000	0

(a) Veja se este homem de negócios deve fazer o investimento.

(b) Vamos supor que um geólogo especialista em poços de petróleo dê sua opinião sobre a produtividade do poço. Se o geólogo sabe que acertará em 60 por cento dos casos e se o homem de negócios segue seu conselho, deverá investir?

(c) Quanto da informação obtida em (b) é válida para o homem de negócios?

9 — Certa empresa não sabe se vai comprar e distribuir 50 000 ou 100 000 unidades de um novo produto. Experiência anterior com casos semelhantes mostra que em 40 por cento das vezes há demanda suficiente para justificar a compra de um maior número de unidades. Os lucros para as duas possibilidades, grande e pequena demanda, aparecem na seguinte tabela **payoff**. (a) Determine a decisão correta para a firma. (b) Vamos supor que uma firma de consultoria de análise de mercado faça um estudo detalhado deste produto particular e de seu potencial provável de vendas, e com base neste estudo diga à empresa se deve comprar o número de unidades pequeno ou grande. Qual seria o grau de fidedignidade deste conselho, se no passado sua opinião esteve correta 70 por cento das vezes?

Demanda	50 000 Unidades	100 000 Unidades
Grande	20 000	50 000
Pequena	10 000	—20 000

Seção 4

10 — Com relação ao problema 7, vamos supor que se tenha feito um inquérito amostral em várias lojas de discos para se determinar a vendagem desse disco específico. Seja **S** os resultados da amostra. Suponhamos que os cálculos baseados na distribuição binomial dêem as probabilidades condicionais.

$$P\{S|0,4\} = 0,10,\ P\{S|0,5\} = 0,10,\ \text{e}\ P\{S|0,6\} = 0,05$$

(a) Use estes valores e a fórmula de Bayes para calcular as probabilidades posteriores $P\{0,4 \mid S\}$, $P\{0,5 \mid S\}$ e $P\{0,6 \mid S\}$.

(b) Usando os resultados de (a), calcule o lucro esperado e determine se a companhia ainda empresa o disco.

11 — Uma companhia especializada na criação e desenvolvimento de novas idéias para brinquedos chegou à concepção de um tipo novo de brinquedo, cuja elaboração está agora considerando. A companhia pode vender os direitos autorais de tal brinquedo ou fabricá-lo. A administração julga que pode vender os direitos por $ 100.000 com um bônus adicional de $ 100.000 se o brinquedo for um sucesso. Mas se a própria companhia fabricar o brinquedo, julga que pode ganhar $ 500.000 se o brinquedo for sucesso, mas perderá $ 200.000 se não for. Por simplicidade vamos considerar o brinquedo como sucesso se obter, pelo menos, certa percentagem do mercado e como insucesso se conseguir menos que esta percentagem. Esta supersimplificação do problema é feita por conveniência de cálculo.

Se a companhia estiver interessada em maximizar lucros, que deveria fazer se (a) a administração está apostando 7 em 10 que o brinquedo será sucesso; (b) fez-se uma pesquisa de mercado e descobriu-se que a probabilidade de se obter o resultado da pesquisa é de 0,04 se o brinquedo for sucesso e de 0,03 se não for sucesso?

	Ação	
Evento	Venda de direitos	Fabricação
Sucesso	200.000	500.000
Insucesso	100.000	—300.000

APÊNDICE

Tabelas

TABELA I — Quadrados e Raízes quadradas.

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
1,00	1,0000	1,00000	3,16228
1,01	1,0201	1,00499	3,17805
1,02	1,0404	1,00995	3,19374
1,03	1,0609	1,01489	3,20936
1,04	1,0816	1,01980	3,22490
1,05	1,1025	1,02470	3,24037
1,06	1,1236	1,02956	3,25576
1,07	1,1449	1,03441	3,27109
1,08	1,1664	1,03923	3,28634
1,09	1,1881	1,04403	3,30151
1,10	1,2100	1,04881	3,31662
1,11	1,2321	1,05357	3,33167
1,12	1,2544	1,05830	3,34664
1,13	1,2769	1,06301	3,36155
1,14	1,2996	1,06771	3,37639
1,15	1,3225	1,07238	3,39116
1,16	1,3456	1,07703	3,40588
1,17	1,3689	1,08167	3,42053
1,18	1,3924	1,08628	3,43511
1,19	1,4161	1,09087	3,44964
1,20	1,4400	1,09545	3,46410
1,21	1,4641	1,10000	3,47851
1,22	1,4884	1,10454	3,49285
1,23	1,5129	1,10905	3,50714
1,24	1,5376	1,11355	3,52136
1,25	1,5625	1,11803	3,53553
1,26	1,5876	1,12250	3,54965
1,27	1,6129	1,12694	3,56371
1,28	1,6384	1,13137	3,57771
1,29	1,6641	1,13578	3,59166
1,30	1,6900	1,14018	3,60555
1,31	1,7161	1,14455	3,61939
1,32	1,7424	1,14891	3,63318
1,33	1,7689	1,15326	3,64692
1,34	1,7956	1,15758	3,66060
1,35	1,8225	1,16190	3,67423
1,36	1,8496	1,16619	3,68782
1,37	1,8769	1,17047	3,70135
1,38	1,9044	1,17473	3,71484
1,39	1,9321	1,17898	3,72827
1,40	1,9600	1,18322	3,74166
1,41	1,9881	1,18743	3,75500
1,42	2,0164	1,19164	3,76829
1,43	2,0449	1,19583	3,78153
1,44	2,0736	1,20000	3,79473
1,45	2,1025	1,20416	3,80789
1,46	2,1316	1,20830	3,82099
1,47	2,1609	1,21244	3,83406
1,48	2,1904	1,21655	3,84708
1,49	2,2201	1,22066	3,86005
1,50	2,2500	1,22474	3,87298

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
1,50	2,2500	1,22474	3,87298
1,51	2,2801	1,22882	3,88587
1,52	2,3104	1,23288	3,89872
1,53	2,3409	1,23693	3,91152
1,54	2,3716	1,24097	3,92428
1,55	2,4025	1,24499	3,93700
1,56	2,4336	1,24900	3,94968
1,57	2,4649	1,25300	3,96232
1,58	2,4964	1,25698	3,97492
1,59	2,5281	1,26095	3,98748
1,60	2,5600	1,26491	4,00000
1,61	2,5921	1,26886	4,01248
1,62	2,6244	1,27279	4,02492
1,63	2,6569	1,27671	4,03733
1,64	2,6896	1,28062	4,04969
1,65	2,7225	1,28452	4,06202
1,66	2,7556	1,28841	4,07431
1,67	2,7889	1,29228	4,08656
1,68	2,8224	1,29615	4,09878
1,69	2,8561	1,30000	4,11096
1,70	2,8900	1,30384	4,12311
1,71	2,9241	1,30767	4,13521
1,72	2,9584	1,31149	4,14729
1,73	2,9929	1,31529	4,15933
1,74	3,0276	1,31909	4,17133
1,75	3,0625	1,32288	4,18330
1,76	3,0976	1,32665	4,19524
1,77	3,1329	1,33041	4,20714
1,78	3,1684	1,33417	4,21900
1,79	3,2041	1,33791	4,23084
1,80	3,2400	1,34164	4,24264
1,81	3,2761	1,34536	4,25441
1,82	3,3124	1,34907	4,26615
1,83	3,3489	1,35277	4,27785
1,84	3,3856	1,35647	4,28952
1,85	3,4225	1,36015	4,30116
1,86	3,4596	1,36382	4,31277
1,87	3,4969	1,36748	4,32435
1,88	3,5344	1,37113	4,33590
1,89	3,5721	1,37477	4,34741
1,90	3,6100	1,37840	4,35890
1,91	3,6481	1,38203	4,37035
1,92	3,6864	1,38564	4,38178
1,93	3,7249	1,38924	4,39318
1,94	3,7636	1,39284	4,40454
1,95	3,8025	1,39642	4,41588
1,96	3,8416	1,40000	4,42719
1,97	3,8809	1,40357	4,43847
1,98	3,9204	1,40712	4,44972
1,99	3,9601	1,41067	4,46094
2,00	4,0000	1,41421	4,47214

N	N^2	$\sqrt{N}$	$\sqrt{10N}$

TABELA I — (continuação)

N	N²	√N	√10N
2,00	4,0000	1,41421	4,47214
2,01	4,0401	1,41774	4,48330
2,02	4,0804	1,42127	4,49444
2,03	4,1209	1,42478	4,50555
2,04	4,1616	1,42829	4,51664
2,05	4,2025	1,43178	4,52769
2,06	4,2436	1,43527	4,53872
2,07	4,2849	1,43875	4,54973
2,08	4,3264	1,44222	4,56070
2,09	4,3681	1,44568	4,57165
2,10	4,4100	1,44914	4,58258
2,11	4,4521	1,45258	4,59347
2,12	4,4944	1,45602	4,60435
2,13	4,5369	1,45945	4,61519
2,14	4,5796	1,46287	4,62601
2,15	4,6225	1,46629	4,63681
2,16	4,6656	1,46969	4,64758
2,17	4,7089	1,47309	4,65833
2,18	4,7524	1,47648	4,66905
2,19	4,7961	1,47986	4.67974
2,20	4,8400	1,48324	4,69042
2,21	4,8841	1,48661	4,70106
2,22	5,9284	1,48997	4,71169
2,23	4,9729	1,49332	4,72229
2,24	5,0176	1,49666	4,73286
2,25	5,0625	1,50000	4,74342
2,26	5,1076	1,50333	4,75395
2,27	5,1529	1,50665	4,76445
2,28	5,1984	1,50997	4,77493
2,29	5,2441	1,51327	4,78539
2,30	5,2900	1,51658	4,79583
2,31	5,3361	1,51987	4,80625
2,32	5,3824	1,52315	4,81664
2,33	5,4289	1,52643	4,82701
2,34	5,4756	1,52971	4,83735
2,35	5,5225	1,53297	4,84768
2,36	5,5696	1,53623	4,85798
2,37	5,6169	1,53948	4,86826
2,38	5,6644	1,54272	4,87852
2,39	5,7121	1,54596	4,88876
2,40	5,7600	1,54919	4,89898
2,41	5,8081	1,55252	4,90918
2,42	5,8564	1,55563	4,91935
2,43	5,9049	1,55885	4,92950
2,44	5,9536	1,56205	4,93964
2,45	6,0025	1,56525	4,94975
2,46	6,0516	1,56844	4,95984
2,47	6,1009	1,57162	4,96991
2,48	6,1054	1,57480	4,97996
2,49	6,2001	1,57797	4,98999
2,50	6,2500	1,58114	5,00000
N	N²	√N	√10N

N	N²	√N	√10N
2,50	6,2500	1,58114	5,00000
2,51	6,3001	1,58430	5,00999
2,52	6,3504	1,58745	5,01996
2,53	6,4009	1,59060	5,02991
2,54	6,4516	1,59374	5,03984
2,55	6,5025	1,59687	5,04975
2,56	6,5536	1,60000	5,05964
2,57	6,6049	1,60312	5,06952
2,58	6,6564	1,60624	5,07937
2,59	6,7081	1,60935	5,08920
2,60	6,7600	1,61245	5,09902
2.61	6,8121	1,61555	5,10882
2.62	6,8644	1,61864	5,11859
2.63	6,9169	1,62173	5,12835
2.64	6,9696	1,62481	5,13809
2.65	7,0225	1,62788	5,14782
2.66	7,0756	1,63095	5,15752
2.67	7,1289	1,63401	5,16720
2.68	7,1824	1,63707	5,17687
2.69	7,2361	1,64012	5,18652
2,70	7,2900	1,64317	5,19615
2,71	7,3441	1,64621	5,20577
2,72	7,3984	1,64924	5,21536
2,73	7,4529	1,65227	5,22494
2,74	7,5076	1,65529	5,23450
2,75	7,5625	1,65831	5,24404
2,76	7,6176	1,66132	5,25357
2,77	7,6729	1,66433	5,26308
2,78	7,7284	1,66733	5,27257
2,79	7,7841	1,67033	5,28205
2,80	7,8400	1,67332	5,29150
2,81	7,8961	1,67631	5,30094
2,82	7,9524	1,67929	5,31037
2,83	8,0089	1,68226	5,31977
2,84	8,0656	1,68523	5,32917
2,85	8,1225	1,68819	5,33854
2,86	8,1796	1,69115	5,34790
2,87	8,2369	1,69411	5,35724
2,88	8,2944	1,69706	5,36656
2,89	8,3521	1,70000	5,37587
2,90	8,4100	1,70294	5,38516
2,91	8,4681	1,70587	5,39444
2,92	8,5264	1,70880	5,40370
2,93	8,5849	1,71172	5,41295
2,94	8,6436	1,71464	5,42218
2,95	8,7025	1,71756	5,43139
2,96	8,7616	1,72047	5,44059
2,97	8,8209	1,72337	5,44977
2,98	8,8804	1,72627	5,45894
2,99	8,9401	1,72916	5,46809
3,00	9,0000	1,73205	5,47723
N	N²	√N	√10N

TABELA I — (continuação)

N	N²	√N	√10N
3,00	9,0000	1,73205	5,47723
3,01	9,0601	1,73494	5,48635
3,02	9,1204	1,73781	5,49545
3,03	9,1809	1,74069	5,50454
3,04	9,2416	1,74356	5,51362
3,05	9,3025	1,74642	5,52268
3,06	9,3636	1,74929	5,53173
3,07	9,4249	1,75214	5,54076
3,08	9,4864	1,75499	5,54977
3,09	9,5481	1,75784	5,55878
3,10	9,6100	1,76068	5,56776
3,11	9,6721	1,76352	5,57674
3,12	9,7344	1,76635	5,58570
3,13	9,7969	1,76918	5,59464
3,14	9,8596	1,77200	5,60357
3,15	9,9225	1,77482	5,61249
3,16	9,9856	1,77764	5,62139
3,17	10,0489	1,78045	5,63028
3,18	10,1124	1,78326	5,63915
3,19	10,1761	1,78606	5,64801
3,20	10,2400	1,78885	5,65685
3,21	10,3041	1,79165	5,66569
3,22	10,3684	1,79444	5,67450
3,23	10,4329	1,79722	5,68331
3,24	10,4976	1,80000	5,69210
3,25	10,5625	1,80278	5,70088
3,26	10,6276	1,80555	5,70964
3,27	10,6929	1,80831	5,71839
3,28	10,7584	1,81108	5,72713
3,29	10,8241	1,81384	5,73585
3,30	10,8900	1,81659	5,74456
3,31	10,9561	1,81934	5,75326
3,32	10,0224	1,82209	5,76194
3,33	11,0889	1,82483	5,77062
3,34	11,1556	1,82757	5,77927
3,35	11,2225	1,83030	5,78792
3,36	11,2896	1,83303	5,79655
3,37	11,3569	1,83576	5,80517
3,38	11,4244	1,83848	5,81378
3,39	11,4921	1,84120	5,82237
3,40	11,5600	1,84391	5,83095
3,41	11,6281	1,84662	5,83952
3,42	11,6964	1,84932	5,84808
3,43	11,7649	1,85203	5,85662
3,44	11,8336	1,85472	5,86515
3,45	11,9025	1,85742	5,87367
3,46	11,9716	1,86011	5,88218
3,47	12,0409	1,86279	5,89067
3,48	12,1104	1,86548	5,89915
3,49	12,1801	1,86815	5,90762
3,50	12,2500	1,87083	5,91608
N	N²	√N	√10N

N	N²	√N	√10N
3,50	12,2500	1,87083	5,91608
3,51	12,3201	1,87350	5,92453
3,52	12,3904	1,87617	5,93296
3,53	12,4609	1,87883	5,94138
3,54	12,5316	1,88149	5,94979
3,55	12,6025	1,88414	5,95819
3,56	12,6736	1,88680	5,96657
3,57	12,7449	1,88944	5,97495
3,58	12,8164	1,89209	5,98331
3,59	12,8881	1,89473	5,99166
3,60	12,9600	1,89737	6,00000
3,61	13,0321	1,90000	6,00833
3,62	13,1044	1,90263	6,01664
3,63	13,1769	1,90526	6,02495
3,64	13,2496	1,90788	6,03324
3,65	13,3225	1,91050	6,04152
3,66	13,3956	1,91311	6,04949
3,67	13,4689	1,91572	6,05805
3,68	13,5424	1,91833	6,06630
3,69	13,6161	1,92094	6,07454
3,70	13,6900	1,92354	6,08276
3,71	13,7641	1,92614	6,09098
3,72	13,8384	1,92873	6,09918
3,73	13,9129	1,93132	6,10737
3,74	13,9876	1,93391	6,11555
3,75	14,0625	1,93649	6,12372
3,76	14,1376	1,93907	6,13188
3,77	14,2129	1,94165	6,14003
3,78	14,2884	1,94422	6,14817
3,79	14,3641	1,94679	6,15630
3,80	14,4400	1,94936	6,16441
3,81	14,5161	1,95192	6,17252
3,82	14,5924	1,95448	6,18061
3,83	14,6689	1,95704	6,18870
3,84	14,7456	1,95959	6,19677
3,85	14,8225	1,96214	6,20484
3,86	14,8996	1,96469	6,21289
3,87	14,9769	1,96723	6,22093
3,88	15,0544	1,96977	6,22896
3,89	15,1321	1,97231	6,23699
3,90	15,2100	1,97484	6,24500
3,91	15,2881	1,97737	6,25300
3,92	15,3664	1,97990	6,26099
3,93	15,4449	1,98242	6,26897
3,94	15,5236	1,98494	6,27694
3,95	15,6025	1,98746	6,28490
3,96	15,6816	1,98997	6,29285
3,97	15,7609	1,99249	6,30079
3,98	15,8404	1,99499	6,30872
3,99	15,9201	1,99750	6,31644
4,00	16,0000	2,00000	6,32456
N	N²	√N	√10N

TABELA I — (continuação)

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
4,00	16,0000	2,00000	6,32456
4,01	16,0801	2,00250	6,33246
4,02	16,1604	2,00499	6,34035
4,03	16,2409	2,00749	6,34823
4,04	16,3216	2,00998	6,35610
4,05	16,4025	2,01246	6,36396
4,06	16,4836	2,01494	6,37181
4,07	16,5649	2,01742	6,37966
4,08	16,6464	2,01990	6,38749
4,09	16,7281	2,02237	6,39531
4,10	16,8100	2,02485	6,40312
4,11	16,8921	2,02731	6,41093
4,12	16,9744	2,02978	6,41872
4,13	17,0569	2,03224	6,42651
4,14	17,1396	2,03470	6,43428
4,15	17,2225	2,03715	6,44205
4,16	17,3056	2,03961	6,44981
4,17	17,3889	2,04206	6,45755
4,18	17,4724	2,04450	6,46529
4,19	17,5561	2,04695	6,47302
4,20	17,6400	2,04939	6,48074
4,21	17,7241	2,05183	6,48845
4,22	17,8084	2,05426	6,49615
4,23	17,8929	2,05670	6,50384
4,24	17,9776	2,05913	6,51153
4,25	18,0625	2,06155	6,51920
4,26	18,1476	2,06398	6,52687
4,27	18,2329	2,06640	6,53452
4,28	18,3184	2,06882	6,54217
4,29	18,4041	2,07123	6,54981
4,30	18,4900	2,07364	6,55744
4,31	18,5761	2,07605	6,66506
4,32	18,6624	2,07846	6,57267
4,33	18,7489	2,08087	6,58027
4,34	18,8356	2,08327	6,58787
4,35	18,9225	2,08567	6,59545
4,36	19,0096	2,08806	6,60303
4,37	19,0969	2,09045	6,61060
4,38	19,1844	2,09284	6,61816
4,39	19,2721	2,09523	6,62571
4,40	19,3600	2,09762	6,63325
4,41	19,4481	2,10000	6,64078
4,42	19,5364	2,10238	6,64831
4,43	19,6249	2,10476	6,65582
4,44	19,7136	2,10713	6,66333
4,45	19,8025	2,10950	6,67083
4,46	19,8916	2,11187	6,67832
4,47	19,9809	2,11424	6,68581
4,48	20,0704	2,11660	6,69328
4,49	20,1601	2,11896	6,70075
4,50	20,2500	2,12132	6,70820
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
4,50	20,2500	2,12132	6,70820
4,51	20,3401	2,12368	6,71565
4,52	20,4304	2,12603	6,72309
4,53	20,5209	2,12838	6,73053
4,54	20,6116	2,13073	6,73795
4,55	20,7025	2,13307	6,74537
4,56	20,7936	2,13542	6,75278
4,57	20,8849	2,13776	6,76018
4,58	20,9764	2,14009	6,76757
4,59	21,0681	2,14243	6,77495
4,60	21,1600	2,14476	6,78233
4,61	21,2521	2,14709	6,78970
4,62	21,3444	2,14942	6,79706
4,63	21,4369	2,15174	6,80441
4,64	21,5296	2,15407	6,81175
4,65	21,6225	2,15639	6,81909
4,66	21,7156	2,15870	6,82642
4,67	21,8089	2,16102	6,83374
4,68	21,9024	2,16333	6,84105
4,69	21,9961	2,16564	6,84836
4,70	22,0900	2,16795	6,85565
4,71	22,1841	2,17025	6,86294
4,72	22,2784	2,17256	6,87023
4,73	22,3729	2,17486	6,87750
4,74	22,4676	2,17715	6,88477
4,75	22,5625	2,17945	6,89202
4,76	22,6576	2,18174	6,89928
4,77	22,7529	2,18403	6,90652
4,78	22,8484	2,18632	6,91375
4,79	22,9441	2,18861	6,92098
4,80	23,0400	2,19089	6,92820
4,81	23,1361	2,19317	6,93542
4,82	23,2324	2,19545	6,94262
4,83	23,3289	2,19773	6,94982
4,84	23,4256	2,20000	6,95701
4,85	23,5225	2,20227	6,96419
4,86	23,6196	2,20454	6,97137
4,87	23,7169	2,20681	6,97854
4,88	23,8144	2,20907	6,98570
4,89	23,9121	2,21133	6,99285
4,90	24,0100	2,21359	7,00000
4,91	24,1081	2,21585	7,00714
4,92	24,2064	2,21811	7,01427
4,93	24,3049	2,22036	7,02140
4,94	24,4036	2,22261	7,02851
4,95	24,5025	2,22486	7,03562
4,96	24,6016	2,22711	7,04273
4,97	24,7009	2,22935	7,04982
4,98	24,8004	2,23159	7,05691
4,99	24,9001	2,23383	7,06399
5,00	25,0000	2,23607	7,07107
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

TABELA I — (continuação)

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
5,00	25,0000	2,23607	7,07107
5,01	25,1001	2,23830	7,07814
5,02	25,2004	2,24054	7,08520
5,03	25,3009	2,24277	7,09225
5,04	25,4016	2,24499	7,09930
5,05	25,5025	2,24722	7,10634
5,06	25,6036	2,24944	7,11337
5,07	25,7049	2,25167	7,12039
5,08	25,8064	2,25389	7,12741
5,09	25,9081	2,25610	7,13442
5,10	26,0100	2,25832	7,14143
5,11	26,1121	2,26053	7,14843
5,12	26,2144	2,26274	7,15542
5,13	26,3169	2,26495	7,16240
5,14	26,4196	2,26716	7,16938
5,15	26,5225	2,26936	7,17635
5,16	26,6256	2,27156	7,18331
5,17	26,7289	2,27376	7,19027
5,18	26,8324	2,27596	7,19722
5,19	26,9361	2,27816	7,20417
5,20	27,0400	2,28035	7,21110
5,21	27,1441	2,28254	7,21803
5,22	27,2484	2,28473	7,22496
5,23	27,3529	2,28692	7,23187
5,24	27,4576	2,28910	7,23838
5,25	27,5625	2,29129	7,24569
5,26	27,6676	2,29347	7,25259
5,27	27,7729	2,29565	7,25948
5,28	27,8784	2,29783	7,26636
5,29	27,9841	2,30000	7,27324
5,30	28,0900	2,30217	7,28011
5,31	28,1961	2,30434	7,28697
5,32	28,3024	2,30651	7,29383
5,33	28,4089	2,30868	7,30068
5,34	28,5156	2,31084	7,30753
5,35	28,6225	2,31301	7,31437
5,36	28,7296	2,31517	7,32120
5,37	28,8369	2,31733	7,32803
5,38	28,9444	2,31948	7,33485
5,39	29,0521	2,32164	7,34166
5,40	29,1600	2,32379	7,34847
5,41	29,2681	2,32594	7,35527
5,42	29,3764	2,32809	7,36206
5,43	29,4849	2,33024	7,36885
5,44	29,5936	2,33238	7,37564
5,45	29,7025	2,33452	7,38241
5,46	29,8116	2,33666	7,38918
5,47	29,9209	2,33880	7,39594
5,48	30,0304	2,34094	7,40270
5,49	30,1401	2,34307	7,40945
5,50	30,2500	2,34521	7,41620
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

N	N^2	$\sqrt{N}$	$\sqrt{10N}$
5,50	30,2500	2,34521	7,41620
5,51	30,3601	2,34734	7,42294
5,52	30,4704	2,34947	7,42967
5,53	30,5809	2,35160	7,43640
5,54	30,6916	2,35372	7,44312
5,55	30,8025	2,35584	7,44983
5,56	30,9136	2,35797	7,45654
5,57	31,0249	2,36008	7,46324
5,58	31,1364	2,36220	7,46994
5,59	31,2481	2,36432	7,47663
5,60	31,3600	2,36643	7,48331
5,61	31,4721	2,36854	7,48999
5,62	31,5844	2,37065	7,49667
5,63	31,6969	2,37276	7,50333
5,64	31,8096	2,37487	7,50999
5,65	31,9225	2,37697	7,51665
5,66	32,0356	2,37908	7,52330
5,67	32,1489	2,38118	7,52994
5,68	32,2624	2,38328	7,53658
5,69	32,3761	2,38537	7,54321
5,70	32,4900	2,38747	7,54983
5,71	32,6041	2,38956	7,55645
5,72	32,7184	2,39165	7,56307
5,73	32,8329	2,39374	7,56968
5,74	32,9476	2,39583	7,57628
5,75	33,0625	2,39792	7,58288
5,76	33,1776	2,40000	7,58947
5,77	33,2929	2,40208	7,59605
5,78	33,4084	2,40416	7,60263
5,79	33,5241	2,40624	7,60920
5,80	33,6400	2,40832	7,61577
5,81	33,7561	2,41039	7,62234
5,82	33,8724	2,41247	7,62889
5,83	33,9889	2,41454	7,63544
5,84	34,1056	2,41661	7,64199
5,85	34,2225	2,41868	7,64853
5,86	34,3396	2,42074	7,65506
5,87	34,4569	2,42281	7,66159
5,88	34,5744	2,42487	7,66812
5,89	34,6921	2,42693	7,67463
5,90	34,8100	2,42899	7,68115
5,91	34,9281	2,43105	7,68765
5,92	35,0464	2,43311	7,69415
5,93	35,1649	2,43516	7,70065
5,94	35,2836	2,43721	7,70714
5,95	35,4025	2,43926	7,71362
5,96	35,5216	2,44131	7,72010
5,97	35,6409	2,44336	7,72658
5,98	35,7604	2,44540	7,73305
5,99	35,8801	2,44745	7,73951
6,00	36,0000	2,44949	7,74597
N	N^2	$\sqrt{N}$	$\sqrt{10N}$

TABELA I — (continuação)

N	N²	√N	√10N
6,00	36,0000	2,44949	7,74597
6,01	36,1201	2,45153	7,75242
6,02	36,2404	2,45357	7,75887
6,03	36,3609	2,45561	7,76531
6,04	36,4816	2,45764	7,77174
6,05	36,6025	2,45967	7,77817
6,06	36,7236	2,46171	7,78460
6,07	36,8449	2,46374	7,79102
6,08	36,9664	2,46577	7,79744
6,09	37,0881	2,46779	7,80385
6,10	37,2100	2,46982	7,81025
6,11	37,3321	2,47184	7,81665
6,12	37,4544	2,47386	7,82304
6,13	37,5769	2,47588	7,82943
6,14	37,6996	2,47790	7,83582
6,15	37,8225	2,47992	7,84219
6,16	37,9456	2,48193	7,84857
6,17	38,0689	2,48395	7,85493
6,18	38,1924	2,48596	7,86130
6,19	38,3161	2,48797	7,86766
6,20	38,4400	2,48998	7,87401
6,21	38,5641	2,49199	7,88036
6,22	38,6884	2,49399	7,88670
6,23	38,8129	2,49600	7,89303
6,24	38,9376	2,49800	7,89937
6,25	39,0625	2,50000	7,90569
6,26	39,1876	2,50200	7,91202
6,27	39,3129	2,50400	7,91833
6,28	39,4384	2,50599	7,92465
6,29	39,5641	2,50799	7,93095
6,30	39,6900	2,50998	7,93725
6,31	39,8161	2,51197	7,94355
6,32	39,9424	2,51396	7,94984
6,33	40,0689	2,51595	7,95613
6,34	40,1956	2,51794	7,96241
6,35	40,3225	2,51992	7,96869
6,36	40,4496	2,52190	7,97496
6,37	40,5769	2,52389	7,98123
6,38	40,7044	2,52587	7,98749
6,39	40,8321	2,52784	7,99375
6,40	40,9600	2,52982	8,00000
6,41	41,0881	2,53180	8,00625
6,42	41,2164	2,53377	8,01249
6,43	41,3449	2,53574	8,01873
6,44	41,4736	2,53772	8,02496
6,45	41,6025	2,53969	8,03119
6,46	41,7316	2,54165	8,03741
6,47	41,8609	2,54362	8,04363
6,48	41,9904	2,54558	8,04984
6,49	42,1201	2,54755	8,05605
6,50	42,2500	2,54951	8,06226
N	N²	√N	√10N

N	N²	√N	√10N
6,50	42,2500	2,54951	8,06226
6,51	42,3801	2,55147	8,06846
6,52	42,5104	2,55343	8,07465
6,53	42,6409	2,55539	8,08084
6,54	42,7716	2,55734	8,08703
6,55	42,9025	2,55930	8,09321
6,56	43,0336	2,56125	8,09938
6,57	43,1649	2,56320	8,10555
6,58	43,2964	2,56515	8,11172
6,59	43,4281	2,56710	8,11788
6,60	43,5600	2,56905	8,12404
6,61	43,6921	2,57099	8,13019
6,62	43,8244	2,57294	8,13634
6,63	43,9569	2,57488	8,14248
6,64	44,0896	2,57682	8,14862
6,65	44,2225	2,57876	8,15475
6,66	44,3556	2,58070	8,16088
6,67	44,4889	2,58263	8,16701
6,68	44,6224	2,58457	8,17313
6,69	44,7561	2,58650	8,17924
6,70	44,8900	2,58844	8,18535
6,71	45,0241	2,59037	8,19146
6,72	45,1584	2,59230	8,19756
6,73	45,2929	2,59422	8,20366
6,74	45,4276	2,59615	8,20975
6,75	45,5625	2,59808	8,21584
6,76	45,6976	2,60000	8,22192
6,77	45,8329	2,60192	8,22800
6,78	45,9684	2,60384	8,23408
6,79	46,1041	2,60576	8,24015
6,80	46,2400	2,60768	8,24621
6,81	46,3761	2,60960	8,25227
6,82	46,5124	2,61151	8,25833
6,83	46,6489	2,61343	8,26438
6,84	46,7856	2,61534	8,27043
6,85	46,9225	2,61725	8,27647
6,86	47,0596	2,61916	8,28251
6,87	47,1969	2,62107	8,28855
6,88	47,3344	2,62298	8,29458
6,89	47,4721	2,62488	8,30060
6,90	47,6100	2,62679	8,30662
6,91	47,7481	2,62869	8,31264
6,92	47,8864	2,63059	8,31865
6,93	48,0249	2,63249	8,32466
6,94	48,1636	2,63439	8,33067
6,95	48,3025	2,63629	8,33667
6,96	48,4416	2,63818	8,34266
6,97	48,5809	2,64008	8,34865
6,98	48,7204	2,64197	8,35464
6,99	48,8601	2,64386	8,36062
7,00	49,0000	2,64575	8,36660
N	N²	√N	√10N

TABELA I — (continuação)

N	N²	√N	√10N
7,00	49,0000	2,64575	8,36660
7,01	49,1401	2,64764	8,37257
7,02	49,2804	2,64953	8,37854
7,03	49,4209	2,65141	8,38451
7,04	49,5616	2,65330	8,39047
7,05	49,7025	2,65518	8,39643
7,06	49,8436	2,65707	8,40238
7,07	49,9849	2,65895	8,40833
7,08	50,1264	2,66083	8,41427
7,09	50,2681	2,66271	8,42021
7,10	50,4100	2,66458	8,42615
7,11	50,5521	2,66646	8,43208
7,12	50,6944	2,66833	8,43801
7,13	50,8369	2,67021	8,44393
7,14	50,9796	2,67208	8,44985
7,15	51,1225	2,67395	8,45577
7,16	51,2656	2,67582	8,46168
7,17	51,4089	2,67769	8,46759
7,18	51,5524	2,67955	8,47349
7,19	51,6961	2,68142	8,47939
7,20	51,8400	2,68328	8,48528
7,21	51,9841	2,68514	8,49117
7,22	52,1284	2,68701	8,49706
7,23	52,2729	2,68887	8,50294
7,24	52,4176	2,69072	8,50882
7,25	52,5625	2,69258	8,51469
7,26	52,7076	2,69444	8,52056
7,27	52,8529	2,69629	8,52643
7,28	52,9984	2,69815	8,53229
7,29	53,1441	2,70000	8,53815
7,30	53,2900	2,70185	8,54400
7,31	53,4361	2,70370	8,54985
7,32	53,5824	2,70555	8,55570
7,33	53,7289	2,70740	8,56154
7,34	53,8756	2,70924	8,56738
7,35	54,0225	2,71109	8,57321
7,36	54,1696	2,71293	8,57904
7,37	54,3169	2,71477	8,58487
7,38	54,4644	2,71662	8,59069
7,39	54,6121	2,71846	8,59651
7,40	54,7600	2,72029	8,60233
7,41	54,9081	2,72213	8,60814
7,42	55,0564	2,72397	8,61394
7,43	55,2049	2,72580	8,61974
7,44	55,3536	2,72764	8,62554
7,45	55,5025	2,72947	8,63134
7,46	55,6516	2,73130	8,63713
7,47	55,8009	2,73313	8,64292
7,48	55,9504	2,73496	8,64870
7,49	56,1001	2,73679	8,65448
7,50	56,2500	2,73861	8,66025
N	N²	√N	√10N

N	N²	√N	√10N
7,50	56,2500	2,73861	8,66025
7,51	56,4001	2,74044	8,66603
7,52	56,5504	2,74226	8,67179
7,53	56,7009	2,74408	8,67756
7,54	56,8516	2,74591	8,68332
7,55	57,0025	2,74773	8,68907
7,56	57,1536	2,74955	8,69483
7,57	57,3049	2,75136	8,70057
7,58	57,4564	2,75318	8,70632
7,59	57,6081	2,75500	8,71206
7,60	57,7600	2,75681	8,71780
7,61	57,9121	2,75862	8,72353
7,62	58,0644	2,76043	8,72926
7,63	58,2169	2,76225	8,73499
7,64	58,3696	2,76405	8,74071
7,65	58,5225	2,76586	8,74643
7,66	58,6756	2,76767	8,75214
7,67	58,8289	2,76948	8,75785
7,68	58,9824	2,77128	8,76356
7,69	59,1361	2,77308	8,76926
7,70	59,2900	2,77489	8,77496
7,71	59,4441	2,77669	8,78066
7,72	59,5984	2,77849	8,78635
7,73	59,7529	2,78029	8,79204
7,74	59,9076	2,78209	8,79773
7,75	60,0625	2,78388	8,80341
7,76	60,2176	2,78568	8,80909
7,77	60,3729	2,78747	8,81476
7,78	60,5284	2,78927	8,82043
7,79	60,6841	2,79106	8,82610
7,80	60,8400	2,79285	8,83176
7,81	60,9961	2,79464	8,83742
7,82	61,1524	2,79643	8,84308
7,83	61,3089	2,79821	8,84873
7,84	61,4656	2,80000	8,85438
7,85	61,6225	2,80179	8,86002
7,86	61,7796	2,80357	8,86566
7,87	61,9369	2,80535	8,87130
7,88	62,0944	2,80713	8,87694
7,89	62,2521	2,80891	8,88257
7,90	62,4100	2,81069	8,88819
7,91	62,5681	2,81247	8,89382
7,92	62,7264	2,81425	8,89944
7,93	62,8849	2,81603	8,90505
7,94	63,0436	2,81780	8,91067
7,95	63,2025	2,81957	8,91628
7,96	63,3616	2,82135	8,92188
7,97	63,5209	2,82312	8,92749
7,98	63,6804	2,82489	8,93308
7,99	63,8401	2,82666	8,93868
8,00	64,0000	2,82843	8,94427
N	N²	√N	√10N

TABELA I — (continuação)

N	N²	√N	√10N
8,00	64,0000	2,82843	8,94427
8,01	64,1601	2,83019	8,94986
8,02	64,3204	2,83196	8,95545
8,03	64,4809	2,83373	8,96103
8,04	64,6416	2,83549	8,96660
8,05	64,8025	2,83725	8,97218
8,06	64,9636	2,83901	8,97775
8,07	65,1249	2,84077	8,98332
8,08	65,2864	2,84253	8,98888
8,09	65,4481	2,84429	8,99444
8,10	65,6100	2,84605	9,00000
8,11	65,7721	2,84781	9,00555
8,12	65,9344	2,84956	9,01110
8,13	66,0969	2,85132	9,01665
8,14	66,2596	2,85307	9,02219
8,15	66,4225	2,85482	9,02774
8,16	66,5856	2,85657	9,03327
8,17	66,7489	2,85832	9,03881
8,18	66,9124	2,86007	9,04434
8,19	67,0761	2,86182	9,04986
8,20	67,2400	2,86356	9,05539
8,21	67,4041	2,86531	9,06091
8,22	67,5684	2,86705	9,06642
8,23	67,7329	2,86880	9,07193
8,24	67,8976	2,87054	9,07744
8,25	68,0625	2,87228	9,08295
8,26	68,2276	2,87402	9,08845
8,27	68,3929	2,87576	9,09395
8,28	68,5584	2,87750	9,09945
8,29	68,7241	2,87924	9,10494
8,30	68,8900	2,88097	9,11045
8,31	69,0561	2,88271	9,11592
8,32	69,2224	2,88444	9,12140
8,33	69,3889	2,88617	9,12688
8,34	69,5556	2,88791	9,13236
8,35	69,7225	2,88964	9,13783
8,36	69,8896	2,89137	9,14330
8,37	70,0569	2,89310	9,14877
8,38	70,2244	2,89482	9,15423
8,39	70,3921	2,89655	9,15969
8,40	70,5600	2,89828	9,16515
8,41	70,7281	2,90000	9,17061
8,42	70,8964	2,90172	9,17606
8,43	71,0649	2,90345	9,18150
8,44	71,2336	2,90517	9,18695
8,45	71,4025	2,90689	9,19239
8,46	71,5716	2,90861	9,19783
8,47	71,7409	2,91033	9,20326
8,48	71,9104	2,91204	9,20869
8,49	72,0801	2,91376	9,21412
8,50	72,2500	2,91548	9,21954
N	N²	√N	√10N

N	N²	√N	√10N
8,50	72,2500	2,91548	9,21954
8,51	72,4201	2,91719	9,22497
8,52	72,5904	2,91890	9,23038
8,53	72,7609	2,92062	9,23580
8,54	72,9316	2,92233	9,24121
8,55	73,1025	2,92404	9,24662
8,56	73,2736	2,92575	9,25203
8,57	73,4449	2,92746	9,25743
8,58	73,6164	2,92916	9,26283
8,59	73,7881	2,93087	9,26823
8,60	73,9600	2,93258	9,27362
8,61	74,1321	2,93428	9,27901
8,62	74,3044	2,93598	9,28440
8,63	74,4769	2,93769	9,28978
8,64	74,6496	2,93939	9,29516
8,65	74,8225	2,94109	9,30054
8,66	74,9956	2,94279	9,30591
8,67	75,1689	2,94449	9,31128
8,68	75,3424	2,94618	9,31665
8,69	75,5161	2,94788	9,32202
8,70	75,6900	2,94958	9,32738
8,71	75,8641	2,95127	9,33274
8,72	76,0384	2,95296	9,33809
8,73	76,2129	2,95466	9,34345
8,74	76,3876	2,95635	9,34880
8,75	76,5625	2,95804	9,35414
8,76	76,7376	2,95973	9,35949
8,77	76,9129	2,96142	9,36483
8,78	77,0884	2,96311	9,37017
8,79	77,2641	2,96479	9,37550
8,80	77,4400	2.96648	9,38083
8,81	77,6161	2,96816	9,38616
8,82	77,7924	2,96985	9,39149
8,83	77,9689	2,97153	9,39681
8,84	78,1456	2,97321	9,40213
8,85	78,3225	2,97489	9,40744
8,86	78,4996	2,97658	9,41276
8,87	78,6769	2,97825	9,41807
8,88	78,8544	2,97993	9,42338
8,89	79,0321	2,98161	9,42868
8,90	79,2100	2,98329	9,43398
8,91	79,3881	2,98496	9,43928
8,92	79,5664	2,98664	9,44458
8,93	79,7449	2,98831	9,44987
8,94	79,9236	2,98998	9,45516
8,95	80,1025	2,99166	9,46044
8,96	80,2816	2,99333	9,46573
8,97	80,4609	2,99500	9,47101
8,98	80,6404	2,99666	9,47629
8,99	80,8201	2,99833	9,48156
9,00	81,0000	3,00000	9,48683
N	N²	√N	√10N

TABELA I — (continuação)

N	N²	√N	√10N
9,00	81,0000	3,00000	9,48683
9,01	81,1801	3,00167	9,49210
9,02	81,3604	3,00333	9,49737
9,03	81,5409	3,00500	9,50263
9,04	81,7216	3,00666	9,50789
9,05	81,9025	3,00832	9,51315
9,06	82,0836	3,00998	9,51840
9,07	82,2649	3,01164	9,52365
9,08	82,4464	3,01330	9,52890
9,09	82,6281	3,01496	9,53415
9,10	82,8100	3,01662	9,53939
9,11	82,9921	3,01828	9,54463
9,12	83,1744	3,01993	9,54987
9,13	83,3569	3,02159	9,55510
9,14	83,5396	3,02324	9,56033
9,15	83,7225	3,02490	9,56556
9,16	83,9056	3,02655	9,57079
9,17	84,0889	3,02820	9,57601
9,18	84,2724	3,02985	9,58123
9,19	84,4561	3,03150	9,58645
9,20	84,6400	3,03315	9,59166
9,21	84,8241	3,03480	9,59687
9,22	85,0084	3,03645	9,60208
9,23	85,1929	3,03809	9,60729
9,24	85,3776	3,03974	9,61249
9,25	85,5625	3,04138	9,61769
9,26	85,7476	3,04302	9,62289
9,27	85,9329	3,04467	9,62808
9,28	86,1184	3,04631	9,63328
9,29	86,3041	3,04795	9,63846
9,30	86,4900	3,04959	9,64365
9,31	86,6761	3,05123	9,64883
9,32	86,8624	3,05287	9,65401
9,33	87,0489	3,05450	9,65919
9,34	87,2356	3,05614	9,66437
9,35	87,4225	3,05778	9,66954
9,36	87,6096	3,05941	9,67471
9,37	87,7969	3,06105	9,67988
9,38	87,9844	3,06268	9,68504
9,39	88,1721	3,06431	9,69020
9,40	88,3600	3,06594	9,69536
9,41	88,5481	3,06757	9,70052
9,42	88,7364	3,06920	9,70567
9,43	88,9249	3,07083	9,71082
9,44	89,1136	3,07246	9,71597
9,45	89,3025	3,07409	9,72111
9,46	89,4916	3,07571	9,72625
9,47	89,6809	3,07734	9,73139
9,48	89,8704	3,07896	9,73653
9,49	90,0601	3,08058	9,74166
9,50	90,2500	3,08221	9,74679
N	N²	√N	√10N

N	N²	√N	√10N
9,50	90,2500	3,08221	9,74679
9,51	90,4401	3,08383	9,75192
9,52	90,6304	3,08545	9,75705
9,53	90,8209	3,08707	9,76217
9,54	91,0116	3,08869	9,76729
9,55	91,2025	3,09031	9,77241
9,56	91,3936	3,09192	9,77753
9,57	91,5849	3,09354	9,78264
9,58	91,7764	3,09516	9,78775
9,59	91,9681	3,09677	9,79285
9,60	92,1600	3,09839	9,79796
9,61	92,3521	3,10000	9,80306
9,62	92,5444	3,10161	9,80816
9,63	92,7369	3,10322	9,81326
9,64	92,9296	3,10483	9,81835
9,65	93,1225	3,10644	9,82344
9,66	93,3156	3,10805	9,82853
9,67	93,5089	3,10966	9,83362
9,68	93,7024	3,11127	9,83870
9,69	93,8961	3,11288	9,84378
9,70	94,0900	3,11448	9,84886
9,71	94,2841	3,11609	9,85393
9,72	94,4784	3,11769	9,85901
9,73	94,6729	3,11929	9,86408
9,74	94,8676	3,12090	9,86914
9,75	95,0625	3,12250	9,87421
9,76	95,2576	3,12410	9,87927
9,77	95,4529	3,12570	9,88433
9,78	95,6484	3,12730	9,88939
9,79	95,8441	3,12890	9,89444
9,80	96,0400	3,13050	9,89949
9,81	96,2361	3,13209	9,90454
9,82	96,4324	3,13369	9,90959
9,83	96,6289	3,13528	9,91464
9,84	96,8256	3,13688	9,91968
9,85	97,0225	3,13847	9,92472
9,86	97,2196	3,14006	9,92974
9,87	97,4169	3,14166	9,93479
9,88	97,6144	3,14325	9,93982
9,89	97,8121	3,14484	9,94485
9,90	98,0100	3,14643	9,94987
9,91	98,2081	3,14802	9,95490
9,92	98,4064	3,14960	9,95992
9,93	98,6049	3,15119	9,96494
9,94	98,8036	3,15278	9,96995
9,95	99,0025	3,15436	9,97497
9,96	99,2016	3,15595	9,97998
9,97	99,4009	3,15753	9,98499
9,98	99,6004	3,15911	9,98999
9,99	99,8001	3,16070	9,99500
10,0	100,000	3,16228	10,0000
N	N²	√N	√10N

TABELA II — Coeficientes binomiais $\frac{n!}{x!(n-x)!}$

n \ x	2	3	4	5	6	7	8	9	10
2	1								
3	3	1							
4	6	4	1						
5	10	10	5	1					
6	15	20	15	6	1				
7	21	35	35	21	7	1			
8	28	56	70	56	28	8	1		
9	36	84	126	126	84	36	9	1	
10	45	120	210	252	210	120	45	10	1
11	55	165	330	462	462	330	165	55	11
12	66	220	495	792	924	792	495	220	66
13	78	286	715	1.287	1.716	1.716	1.287	715	286
14	91	364	1.001	2.002	3.003	3.432	3.003	2.002	1.001
15	105	455	1.365	3.003	5.005	6.435	6.435	5.005	3.003
16	120	560	1.820	4.368	8.008	11.440	12.870	11.440	8.008
17	136	680	2.380	6.188	12.376	19.448	24.310	24.310	19.448
18	153	816	3.060	8.568	18.564	31.824	43.758	48.620	43.758
19	171	969	3.876	11.628	27.132	50.388	75.582	92.378	92.378
20	190	1.140	4.845	15.504	38.760	77.520	125.970	167.960	184.756

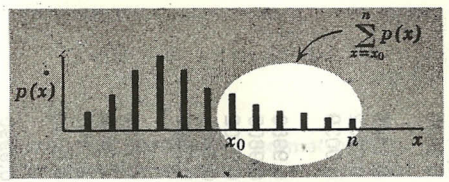


TABELA III — Probabilidades Binomiais Cumulativas na cauda direita.

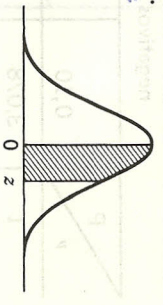
		p									
n	x_0	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
2	1	0,0975	0,1900	0,2775	0,3600	0,4375	0,5100	0,5775	0,6400	0,6975	0,7500
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	1	0,1426	0,2710	0,3859	0,4880	0,5781	0,6570	0,7254	0,7840	0,8336	0,8750
	2	0,0072	0,0280	0,0608	0,1040	0,1562	0,2160	0,2818	0,3520	0,4252	0,5000
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	1	0,1855	0,3439	0,4780	0,5904	0,6836	0,7599	0,8215	0,8704	0,9085	0,9375
	2	0,0140	0,0523	0,1095	0,1808	0,2617	0,3483	0,4370	0,5248	0,6090	0,6875
	3	0,0005	0,0037	0,0120	0,0272	0,0508	0,0837	0,1265	0,1792	0,2415	0,3125
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	1	0,2262	0,4095	0,5563	0,6723	0,7627	0,8319	0,8840	0,9222	0,9497	0,9688
	2	0,0226	0,0815	0,1648	0,2627	0,3672	0,4718	0,5716	0,6630	0,7438	0,8125
	3	0,0012	0,0086	0,0266	0,0579	0,1035	0,1631	0,2352	0,3174	0,4069	0,5000
	4	0,0000	0,0005	0,0022	0,0067	0,0156	0,0308	0,0540	0,0870	0,1312	0,1875
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
6	1	0,2649	0,4686	0,6229	0,7379	0,8220	0,8824	0,9246	0,9533	0,9723	0,9844
	2	0,0328	0,1143	0,2235	0,3447	0,4661	0,5798	0,6809	0,7667	0,8364	0,8906
	3	0,0022	0,0158	0,0473	0,0989	0,1694	0,2557	0,3529	0,4557	0,5585	0,6562
	4	0,0001	0,0013	0,0059	0,0170	0,0376	0,0705	0,1174	0,1792	0,2553	0,3438
	5	0,0000	0,0001	0,0004	0,0016	0,0046	0,0109	0,0223	0,0410	0,0692	0,1094
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	1	0,3017	0,5217	0,6794	0,7903	0,8665	0,9176	0,9510	0,9720	0,9848	0,9922
	2	0,0444	0,1497	0,2834	0,4233	0,5551	0,6706	0,7662	0,8414	0,8976	0,9375
	3	0,0038	0,0257	0,0738	0,1480	0,2436	0,3529	0,4677	0,5801	0,6836	0,7734
	4	0,0002	0,0027	0,0121	0,0333	0,0706	0,1260	0,1998	0,2898	0,3917	0,5000
	5	0,0000	0,0002	0,0012	0,0047	0,0129	0,0288	0,0556	0,0963	0,1529	0,2266
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0038	0,0090	0,0188	0,0357	0,0625
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078

TABELA III — (continuação)

n	x_0	p 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
8	1	0,3366	0,5695	0,7275	0,8322	0,8999	0,9424	0,9681	0,9832	0,9916	0,9961
	2	0,0572	0,1869	0,3428	0,4967	0,6329	0,7447	0,8309	0,8936	0,9368	0,9648
	3	0,0058	0,0381	0,1052	0,2031	0,3215	0,4482	0,5722	0,6846	0,7799	0,8555
	4	0,0004	0,0050	0,0214	0,0563	0,1138	0,1941	0,2936	0,4059	0,5230	0,6367
	5	0,0000	0,0004	0,0029	0,0104	0,0273	0,0580	0,1061	0,1737	0,2604	0,3633
	6	0,0000	0,0000	0,0002	0,0012	0,0042	0,0113	0,0253	0,0498	0,0885	0,1445
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0085	0,0181	0,0352
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	1	0,3698	0,6126	0,7684	0,8658	0,9249	0,9596	0,9793	0,9899	0,9954	0,9980
	2	0,0712	0,2252	0,4005	0,5638	0,6997	0,8040	0,8789	0,9295	0,9615	0,9805
	3	0,0084	0,0530	0,1409	0,2618	0,3993	0,5372	0,6627	0,7682	0,8505	0,9102
	4	0,0006	0,0083	0,0339	0,0856	0,1657	0,2703	0,3911	0,5174	0,6386	0,7461
	5	0,0000	0,0009	0,0056	0,0196	0,0489	0,0988	0,1717	0,2666	0,3786	0,5000
	6	0,0000	0,0001	0,0006	0,0031	0,0100	0,0253	0,0536	0,0994	0,1658	0,2539
	7	0,0000	0,0000	0,0000	0,0003	0,0013	0,0043	0,0112	0,0250	0,0498	0,0898
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0038	0,0091	0,0195
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	1	0,4013	0,6513	0,8031	0,8926	0,9437	0,9718	0,9865	0,9940	0,9975	0,9990
	2	0,0861	0,2639	0,4557	0,6242	0,7560	0,8507	0,9140	0,9536	0,9767	0,9893
	3	0,0115	0,0702	0,1798	0,3222	0,4744	0,6172	0,7384	0,8327	0,9004	0,9453
	4	0,0010	0,0128	0,0500	0,1209	0,2241	0,3504	0,4862	0,6177	0,7340	0,8281
	5	0,0001	0,0016	0,0099	0,0328	0,0781	0,1503	0,2485	0,3669	0,4956	0,6230
	6	0,0000	0,0001	0,0014	0,0064	0,0197	0,0473	0,0949	0,1662	0,2616	0,3770
	7	0,0000	0,0000	0,0001	0,0009	0,0035	0,0106	0,0260	0,0548	0,1020	0,1719
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0048	0,0123	0,0274	0,0547
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0017	0,0045	0,0107
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010

TABELA IV — Áreas de uma distribuição normal-padrão.

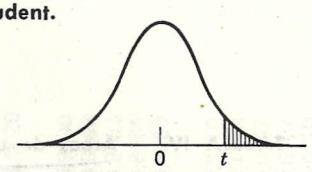
Cada casa da tabela dá a proporção sob a curva total entre $z = 0$ e um valor positivo de z. As áreas para os valores negativos de z são obtidas por simetria.



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4563	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

TABELA V — Distribuição t de Student.

A primeira coluna mostra o número de graus de liberdade (ν). Os cabeçalhos das outras colunas dão as probabilidades (P) de t exceder numericamente o valor da casa. Use simetria para os valores negativos de t.



ν \ P	0,10	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

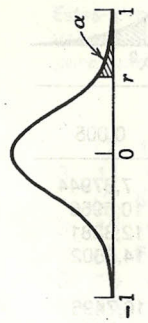


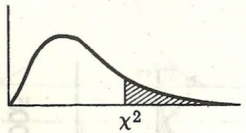
TABELA VI — Valores críticos de r para testar $\rho = 0$.

Para um teste bicaudal, α é duas vezes o valor que aparece no cabeçalho de uma coluna de valores críticos r; dessa forma, para $\alpha = 0{,}05$ escolha a coluna 0,025.

n \ α	0,05	0,025	0,005
5	0,805	0,878	0,959
6	0,729	0,811	0,917
7	0,669	0,754	0,875
8	0,621	0,707	0,834
9	0,582	0,666	0,798
10	0,549	0,632	0,765
11	0,521	0,602	0,735
12	0,497	0,576	0,708
13	0,476	0,553	0,684
14	0,457	0,532	0,661
15	0,441	0,514	0,641
16	0,426	0,497	0,623
17	0,412	0,482	0,606
18	0,400	0,468	0,590
19	0,389	0,456	0,575
20	0,378	0,444	0,561
25	0,337	0,396	0,505
30	0,306	0,361	0,463
35	0,283	0,334	0,430
40	0,264	0,312	0,402
50	0,235	0,279	0,361
60	0,214	0,254	0,330
80	0,185	0,220	0,286
100	0,165	0,196	0,256

TABELA VII — A distribuição χ^2 (qui-quadrado).

A primeira coluna apresenta o número de graus de liberdade (ν). Os cabeçalhos das outras colunas dão as probabilidades (**P**) de χ^2 para exceder o valor da casa. Para $\nu > 100$, tratar $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ como uma variável normal-padrão.



ν \ P	0,995	0,975	0,050	0,025	0,010	0,005
1	$0,0^4 3927$	$0,0^3 9821$	3,84146	5,02389	6,63490	7,87944
2	0,010025	0,050636	5,99147	7,37776	9,21034	10,5966
3	0,071721	0,215795	7,81473	9,34840	11,3449	12,8381
4	0,206990	0,484419	9,48773	11,1433	13,2767	14,8602
5	0,411740	0,831211	11,0705	12,8325	15,0863	16,7496
6	0,675727	1,237347	12,5916	14,4494	16,8119	18,5476
7	0,989265	1,68987	14,0671	16,0128	18,4753	20,2777
8	1,344419	2,17973	15,5073	17,5346	20,0902	21,9550
9	1,734926	2,70039	16,9190	19,0228	21,6660	23,5893
10	2,15585	3,24697	18,3070	20,4831	23,2093	25,1882
11	2,60321	3,81575	19,6751	21,9200	24,7250	26,7569
12	3,07382	4,40379	21,0261	23,3367	26,2170	28,2995
13	3,56503	5,00874	22,3621	24,7356	27,6883	29,8194
14	4,07468	5,62872	23,6848	26,1190	29,1413	31,3193
15	4,60094	6,26214	24,9958	27,4884	30,5779	32,8013
16	5,14224	6,90766	26,2962	28,8454	31,9999	34,2672
17	5,69724	7,56418	27,5871	30,1910	33,4087	35,7185
18	6,26481	8,23075	28,8693	31,5264	34,8053	37,1564
19	6,84398	8,90655	30,1435	32,8523	36,1908	38,5822
20	7,43386	9,59083	31,4104	34,1696	37,5662	39,9968
21	8,03366	10,28293	32,6705	35,4789	38,9321	41,4010
22	8,64272	10,9823	33,9244	36,7807	40,2894	42,7956
23	9,26042	11,6885	35,1725	38,0757	41,6384	44,1813
24	9,88623	12,4001	36,4151	39,3641	42,9798	45,5585
25	10,5197	13,1197	37,6525	40,6465	44,3141	46,9278
26	11,1603	13,8439	38,8852	41,9232	45,6417	48,2899
27	11,8076	14,5733	40,1133	43,1944	46,9630	49,6449
28	12,4613	15,3079	41,3372	44,4607	48,2782	50,9933
29	13,1211	16,0471	42,5569	45,7222	49,5879	52,3356
30	13,7867	16,7908	43,7729	46,9792	50,8922	53,6720
40	20,7065	24,4331	55,7585	59,3417	63,6907	66,7659
50	27,9907	32,3574	67,5048	71,4202	76,1539	79,4900
60	35,5346	40,4817	79,0819	83,2976	88,3794	91,9517
70	43,2752	48,7576	90,5312	95,0231	100,425	104,215
80	51,1720	57,1532	101,879	106,629	112,329	116,321
90	59,1963	65,6466	113,145	118,136	124,116	128,299
100	67,3276	74,2219	124,342	129,561	135,807	140,169

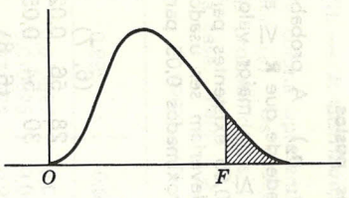
TABELA VIII — Valores críticos da soma dos postos.

Os tamanhos de amostra aparecem entre parênteses (n_1, n_2). A probabilidade associada ao par de valores críticos é a probabilidade de que **R** $\leq$ ao menor valor, ou, então, é a probabilidade de que **R** $\geq$ ao maior valor. Estas probabilidades são as mais próximas de 0,025 e 0,05 existentes para valores inteiros de **R**. Os valores aproximados 0,025 deveriam ser usados para um teste bilateral com $\alpha = 0,05$, e os valores aproximados 0,05 para um teste unilateral.

(2, 4)	(4, 4)	(6, 7)
3 11 0,067	11 25 0,029	28 56 0,026
(2, 5)	12 24 0,057	30 54 0,051
3 13 0,047	(4, 5)	(6, 8)
(2, 6)	12 28 0,032	29 61 0,021
3 15 0,036	13 27 0,056	32 58 0,054
4 14 0,071	(4, 6)	(6, 9)
(2, 7)	12 32 0,019	31 65 0,025
3 17 0,028	14 30 0,057	33 63 0,044
4 16 0,056	(4, 7)	(6, 10)
(2, 8)	13 35 0,021	33 69 0,028
3 19 0,022	15 33 0,055	35 67 0,047
4 18 0,044	(4, 8)	(7, 7)
(2, 9)	14 38 0,024	37 68 0,027
3 21 0,018	16 36 0,055	39 66 0,049
4 20 0,036	(4, 9)	(7, 8)
(2, 10)	15 41 0,025	39 73 0,027
4 22 0,030	17 39 0,053	41 71 0,047
5 21 0,061	(4, 10)	(7, 9)
(3,3)	16 44 0,026	41 78 0,027
6 15 0,050	18 42 0,053	43 76 0,045
(3, 4)	(5, 5)	(7, 10)
6 18 0,028	18 37 0,028	43 83 0,028
7 17 0,057	19 36 0,048	46 80 0,054
(3, 5)	(5, 6)	(8, 8)
6 21 0,018	19 41 0,026	49 87 0,025
7 20 0,036	20 40 0,041	52 84 0,052
(3, 6)	(5, 7)	(8, 9)
7 23 0,024	20 45 0,024	51 93 0,023
8 22 0,048	22 43 0,053	54 90 0,046
(3, 7)	(5, 8)	(8, 10)
8 25 0,033	21 49 0,023	54 98 0,027
9 24 0,058	23 47 0,047	57 95 0,051
(3, 8)	(5, 9)	(9, 9)
8 28 0,024	22 53 0,021	63 108 0,025
9 27 0,042	25 50 0,056	66 105 0,047
(3, 9)	(5, 10)	(9, 10)
9 30 0,032	24 56 0,028	66 114 0,027
10 29 0,050	26 54 0,050	69 111 0,047
(3, 10)	(6, 6)	(10, 10)
9 33 0,024	26 52 0,021	79 131 0,026
11 31 0,056	28 50 0,047	83 127 0,053

TABELA IX — Distribuição F.

Valores da distribuição de **F**, 5% (tipo claro) e 1% (negrito).



Graus de liberdade (ν_2) denominador	Graus de liberdade para o numerador (ν_1)																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161 **4052**	200 **4999**	216 **5403**	225 **5625**	230 **5764**	234 **5859**	237 **5928**	239 **5981**	241 **6022**	242 **6056**	243 **6082**	244 **6106**	245 **6142**	246 **6169**	248 **6208**	249 **6234**	250 **6258**	251 **6286**	252 **6302**	253 **6323**	253 **6334**	254 **6352**	254 **6361**	254 **6366**
2	18,51 **98,49**	19,00 **99,01**	19,16 **99,17**	19,25 **99,25**	19,30 **99,30**	19,33 **99,33**	19,36 **99,34**	19,37 **99,36**	19,38 **99,38**	19,39 **99,40**	19,40 **99,41**	19,41 **99,42**	19,42 **99,43**	19,43 **99,44**	19,44 **99,45**	19,45 **99,46**	19,46 **99,47**	19,47 **99,48**	19,47 **99,48**	19,48 **99,49**	19,49 **99,49**	19,49 **99,49**	19,50 **99,50**	19,50 **99,50**
3	10,13 **34,12**	9,55 **30,81**	9,28 **29,46**	9,12 **28,71**	9,01 **28,24**	8,94 **27,91**	8,88 **27,67**	8,84 **27,49**	8,81 **27,34**	8,78 **27,23**	8,76 **27,13**	8,74 **27,05**	8,71 **26,92**	8,69 **26,83**	8,66 **26,69**	8,64 **26,60**	8,62 **26,50**	8,60 **26,41**	8,58 **26,30**	8,57 **26,27**	8,56 **26,23**	8,54 **26,18**	8,54 **26,14**	8,53 **26,12**
4	7,71 **21,20**	6,94 **18,00**	6,59 **16,69**	6,39 **15,98**	6,26 **15,52**	6,16 **15,21**	6,09 **14,98**	6,04 **14,80**	6,00 **14,66**	5,96 **14,54**	5,93 **14,45**	5,91 **14,37**	5,87 **14,24**	5,84 **14,15**	5,80 **14,02**	5,77 **13,93**	5,74 **13,83**	5,71 **13,74**	5,70 **13,69**	5,68 **13,61**	5,66 **13,57**	5,65 **13,52**	5,64 **13,48**	5,63 **13,46**
5	6,61 **16,26**	5,79 **13,27**	5,41 **12,06**	5,19 **11,39**	5,05 **10,97**	4,95 **10,67**	4,88 **10,45**	4,82 **10,27**	4,78 **10,15**	4,74 **10,05**	4,70 **9,96**	4,68 **9,89**	4,64 **9,77**	4,60 **9,68**	4,56 **9,55**	4,53 **9,47**	4,50 **9,38**	4,46 **9,29**	4,44 **9,24**	4,42 **9,17**	4,40 **9,13**	4,38 **9,07**	4,37 **9,04**	4,36 **9,02**
6	5,99 **13,74**	5,14 **10,92**	4,76 **9,78**	4,53 **9,15**	4,39 **8,75**	4,28 **8,47**	4,21 **8,26**	4,15 **8,10**	4,10 **7,98**	4,06 **7,87**	4,03 **7,79**	4,00 **7,72**	3,96 **7,60**	3,92 **7,52**	3,87 **7,39**	3,84 **7,31**	3,81 **7,23**	3,77 **7,14**	3,75 **7,09**	3,72 **7,02**	3,71 **6,99**	3,69 **6,94**	3,68 **6,90**	3,67 **6,88**
7	5,59 **12,25**	4,74 **9,55**	4,35 **8,45**	4,12 **7,85**	3,97 **7,46**	3,87 **7,19**	3,79 **7,00**	3,73 **6,84**	3,68 **6,71**	3,63 **6,62**	3,60 **6,54**	3,57 **6,47**	3,52 **6,35**	3,49 **6,27**	3,44 **6,15**	3,41 **6,07**	3,38 **5,98**	3,34 **5,90**	3,32 **5,85**	3,29 **5,78**	3,28 **5,75**	3,25 **5,70**	3,24 **5,67**	3,23 **5,65**
8	5,32 **11,26**	4,46 **8,65**	4,07 **7,59**	3,84 **7,01**	3,69 **6,63**	3,58 **6,37**	3,50 **6,19**	3,44 **6,03**	3,39 **5,91**	3,34 **5,82**	3,31 **5,74**	3,28 **5,67**	3,23 **5,56**	3,20 **5,48**	3,15 **5,36**	3,12 **5,28**	3,08 **5,20**	3,05 **5,11**	3,03 **5,06**	3,00 **5,00**	2,98 **4,96**	2,96 **4,91**	2,94 **4,88**	2,93 **4,86**
9	5,12 **10,56**	4,26 **8,02**	3,86 **6,99**	3,63 **6,42**	3,48 **6,06**	3,37 **5,80**	3,29 **5,62**	3,23 **5,47**	3,18 **5,35**	3,13 **5,26**	3,10 **5,18**	3,07 **5,11**	3,02 **5,00**	2,98 **4,92**	2,93 **4,80**	2,90 **4,73**	2,86 **4,64**	2,82 **4,56**	2,80 **4,51**	2,77 **4,45**	2,76 **4,41**	2,73 **4,36**	2,72 **4,33**	2,71 **4,31**
10	4,96 **10,04**	4,10 **7,56**	3,71 **6,55**	3,48 **5,99**	3,33 **5,64**	3,22 **5,39**	3,14 **5,21**	3,07 **5,06**	3,02 **4,95**	2,97 **4,85**	2,94 **4,78**	2,91 **4,71**	2,86 **4,60**	2,82 **4,52**	2,77 **4,41**	2,74 **4,33**	2,70 **4,25**	2,67 **4,17**	2,64 **4,12**	2,61 **4,05**	2,59 **4,01**	2,56 **3,96**	2,55 **3,93**	2,54 **3,91**
11	4,84 **9,65**	3,98 **7,20**	3,59 **6,22**	3,36 **5,67**	3,20 **5,32**	3,09 **5,07**	3,01 **4,88**	2,95 **4,74**	2,90 **4,63**	2,86 **4,54**	2,82 **4,46**	2,79 **4,40**	2,74 **4,29**	2,70 **4,21**	2,65 **4,10**	2,61 **4,02**	2,57 **3,94**	2,53 **3,86**	2,50 **3,80**	2,47 **3,74**	2,45 **3,70**	2,42 **3,66**	2,41 **3,62**	2,40 **3,60**
12	4,75 **9,33**	3,88 **6,93**	3,49 **5,95**	3,26 **5,41**	3,11 **5,06**	3,00 **4,82**	2,92 **4,65**	2,85 **4,50**	2,80 **4,39**	2,76 **4,30**	2,72 **4,22**	2,69 **4,16**	2,64 **4,05**	2,60 **3,98**	2,54 **3,86**	2,50 **3,78**	2,46 **3,70**	2,42 **3,61**	2,40 **3,56**	2,36 **3,49**	2,35 **3,46**	2,32 **3,41**	2,31 **3,38**	2,30 **3,36**
13	4,67 **9,07**	3,80 **6,70**	3,41 **5,74**	3,18 **5,20**	3,02 **4,86**	2,92 **4,62**	2,84 **4,44**	2,77 **4,30**	2,72 **4,19**	2,67 **4,10**	2,63 **4,02**	2,60 **3,96**	2,55 **3,85**	2,51 **3,78**	2,46 **3,67**	2,42 **3,59**	2,38 **3,51**	2,34 **3,42**	2,32 **3,37**	2,28 **3,30**	2,26 **3,27**	2,24 **3,21**	2,22 **3,18**	2,21 **3,16**
14	4,60 **8,86**	3,74 **6,51**	3,34 **5,56**	3,11 **5,03**	2,96 **4,69**	2,85 **4,46**	2,77 **4,28**	2,70 **4,14**	2,65 **4,03**	2,60 **3,94**	2,56 **3,86**	2,53 **3,80**	2,48 **3,70**	2,44 **3,62**	2,39 **3,51**	2,35 **3,43**	2,31 **3,34**	2,27 **3,26**	2,24 **3,21**	2,21 **3,14**	2,19 **3,11**	2,16 **3,06**	2,14 **3,02**	2,13 **3,00**
15	4,54 **8,68**	3,68 **6,36**	3,29 **5,42**	3,06 **4,89**	2,90 **4,56**	2,79 **4,32**	2,70 **4,14**	2,64 **4,00**	2,59 **3,89**	2,55 **3,80**	2,51 **3,73**	2,48 **3,67**	2,43 **3,56**	2,39 **3,48**	2,33 **3,36**	2,29 **3,29**	2,25 **3,20**	2,21 **3,12**	2,18 **3,07**	2,15 **3,00**	2,12 **2,97**	2,10 **2,92**	2,08 **2,89**	2,07 **2,87**
16	4,49 **8,53**	3,63 **6,23**	3,24 **5,29**	3,01 **4,77**	2,85 **4,44**	2,74 **4,20**	2,66 **4,03**	2,59 **3,89**	2,54 **3,78**	2,49 **3,69**	2,45 **3,61**	2,42 **3,55**	2,37 **3,45**	2,33 **3,37**	2,28 **3,25**	2,24 **3,18**	2,20 **3,10**	2,16 **3,01**	2,13 **2,96**	2,09 **2,89**	2,07 **2,86**	2,04 **2,80**	2,02 **2,77**	2,01 **2,75**
17	4,45 **8,40**	3,59 **6,11**	3,20 **5,18**	2,96 **4,67**	2,81 **4,34**	2,70 **4,10**	2,62 **3,93**	2,55 **3,79**	2,50 **3,68**	2,45 **3,59**	2,41 **3,52**	2,38 **3,45**	2,33 **3,35**	2,29 **3,27**	2,23 **3,16**	2,19 **3,08**	2,15 **3,00**	2,11 **2,92**	2,08 **2,86**	2,04 **2,79**	2,02 **2,76**	1,99 **2,70**	1,97 **2,67**	1,96 **2,65**
18	4,41 **8,28**	3,55 **6,01**	3,16 **5,09**	2,93 **4,58**	2,77 **4,25**	2,66 **4,01**	2,58 **3,85**	2,51 **3,71**	2,46 **3,60**	2,41 **3,51**	2,37 **3,44**	2,34 **3,37**	2,29 **3,27**	2,25 **3,19**	2,19 **3,07**	2,15 **3,00**	2,11 **2,91**	2,07 **2,83**	2,04 **2,78**	2,00 **2,71**	1,98 **2,68**	1,95 **2,62**	1,93 **2,59**	1,92 **2,57**
19	4,38 **8,18**	3,52 **5,93**	3,13 **5,01**	2,90 **4,50**	2,74 **4,17**	2,63 **3,94**	2,55 **3,77**	2,48 **3,63**	2,43 **3,52**	2,38 **3,43**	2,34 **3,36**	2,31 **3,30**	2,26 **3,19**	2,21 **3,12**	2,15 **3,00**	2,11 **2,92**	2,07 **2,84**	2,02 **2,76**	2,00 **2,70**	1,96 **2,63**	1,94 **2,60**	1,91 **2,54**	1,90 **2,51**	1,88 **2,49**
20	4,35 **8,10**	3,49 **5,85**	3,10 **4,94**	2,87 **4,43**	2,71 **4,10**	2,60 **3,87**	2,52 **3,71**	2,45 **3,56**	2,40 **3,45**	2,35 **3,37**	2,31 **3,30**	2,28 **3,23**	2,23 **3,13**	2,18 **3,05**	2,12 **2,94**	2,08 **2,86**	2,04 **2,77**	1,99 **2,69**	1,96 **2,63**	1,92 **2,56**	1,90 **2,53**	1,87 **2,47**	1,85 **2,44**	1,84 **2,42**
21	4,32 **8,02**	3,47 **5,78**	3,07 **4,87**	2,84 **4,37**	2,68 **4,04**	2,57 **3,81**	2,49 **3,65**	2,42 **3,51**	2,37 **3,40**	2,32 **3,31**	2,28 **3,24**	2,25 **3,17**	2,20 **3,07**	2,15 **2,99**	2,09 **2,88**	2,05 **2,80**	2,00 **2,72**	1,96 **2,63**	1,93 **2,58**	1,89 **2,51**	1,87 **2,47**	1,84 **2,42**	1,82 **2,38**	1,81 **2,36**
22	4,30 **7,94**	3,44 **5,72**	3,05 **4,82**	2,82 **4,31**	2,66 **3,99**	2,55 **3,76**	2,47 **3,59**	2,40 **3,45**	2,35 **3,35**	2,30 **3,26**	2,26 **3,18**	2,23 **3,12**	2,18 **3,02**	2,13 **2,94**	2,07 **2,83**	2,03 **2,75**	1,98 **2,67**	1,93 **2,58**	1,91 **2,53**	1,87 **2,46**	1,84 **2,42**	1,81 **2,37**	1,80 **2,33**	1,78 **2,31**
23	4,28 **7,88**	3,42 **5,66**	3,03 **4,76**	2,80 **4,26**	2,64 **3,94**	2,53 **3,71**	2,45 **3,54**	2,38 **3,41**	2,32 **3,30**	2,28 **3,21**	2,24 **3,14**	2,20 **3,07**	2,14 **2,97**	2,10 **2,89**	2,04 **2,78**	2,00 **2,70**	1,96 **2,62**	1,91 **2,53**	1,88 **2,48**	1,84 **2,41**	1,82 **2,37**	1,79 **2,32**	1,77 **2,28**	1,76 **2,26**
24	4,26 **7,82**	3,40 **5,61**	3,01 **4,72**	2,78 **4,22**	2,62 **3,90**	2,51 **3,67**	2,43 **3,50**	2,36 **3,36**	2,30 **3,25**	2,26 **3,17**	2,22 **3,09**	2,18 **3,03**	2,13 **2,93**	2,09 **2,85**	2,02 **2,74**	1,98 **2,66**	1,94 **2,58**	1,89 **2,49**	1,86 **2,44**	1,82 **2,36**	1,80 **2,33**	1,76 **2,27**	1,74 **2,23**	1,73 **2,21**
25	4,24 **7,77**	3,38 **5,57**	2,99 **4,68**	2,76 **4,18**	2,60 **3,86**	2,49 **3,63**	2,41 **3,46**	2,34 **3,32**	2,28 **3,21**	2,24 **3,13**	2,20 **3,05**	2,16 **2,99**	2,11 **2,89**	2,06 **2,81**	2,00 **2,70**	1,96 **2,62**	1,92 **2,54**	1,87 **2,45**	1,84 **2,40**	1,80 **2,32**	1,77 **2,29**	1,74 **2,23**	1,72 **2,19**	1,71 **2,17**

TABELA IX — (continuação)

Graus de liberdade (ν_2) denominador	Graus de liberdade para o numerador (ν_1)																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
26	4,22 **7,72**	3,37 **5,53**	2,89 **4,64**	2,74 **4,14**	2,59 **3,82**	2,47 **3,59**	2,39 **3,42**	2,32 **3,29**	2,27 **3,17**	2,22 **3,09**	2,18 **3,02**	2,15 **2,96**	2,10 **2,86**	2,05 **2,77**	1,99 **2,66**	1,95 **2,58**	1,90 **2,50**	1,85 **2,41**	1,82 **2,36**	1,78 **2,28**	1,76 **2,25**	1,72 **2,19**	1,70 **2,15**	1,69 **2,13**
27	4,21 **7,68**	3,35 **5,49**	2,96 **4,60**	2,73 **4,11**	2,57 **3,79**	2,46 **3,56**	2,37 **3,39**	2,30 **3,26**	2,25 **3,14**	2,20 **3,06**	2,16 **2,98**	2,13 **2,93**	2,08 **2,83**	2,03 **2,74**	1,97 **2,63**	1,93 **2,55**	1,88 **2,47**	1,84 **2,38**	1,80 **2,33**	1,76 **2,25**	1,74 **2,21**	1,71 **2,16**	1,68 **2,12**	1,67 **2,10**
28	4,20 **7,64**	3,34 **5,45**	2,95 **4,57**	2,71 **4,07**	2,56 **3,76**	2,44 **3,53**	2,36 **3,36**	2,29 **3,23**	3,24 **3,11**	2,19 **3,03**	2,15 **2,95**	2,12 **2,90**	2,06 **2,80**	2,02 **2,71**	1,96 **2,60**	1,91 **2,52**	1,87 **2,44**	1,81 **2,35**	1,78 **2,30**	1,75 **2,22**	1,72 **2,18**	1,69 **2,13**	1,67 **2,09**	1,65 **2,06**
29	4,18 **7,60**	3,33 **5,52**	2,93 **4,54**	2,70 **4,04**	2,54 **3,73**	2,43 **3,50**	2,35 **3,33**	2,28 **3,20**	2,22 **3,08**	2,18 **3,00**	2,14 **2,92**	2,10 **2,87**	2,05 **2,77**	2,00 **2,68**	1,94 **2,57**	1,90 **2,49**	1,85 **2,41**	1,80 **2,32**	1,77 **2,27**	1,73 **2,19**	1,71 **2,15**	1,68 **2,10**	1,65 **2,06**	1,64 **2,03**
30	4,17 **7,56**	3,32 **5,39**	2,92 **4,51**	2,69 **4,02**	2,53 **3,70**	2,42 **3,47**	2,34 **3,30**	2,27 **3,17**	2,21 **3,06**	2,16 **2,98**	2,12 **2,00**	2,09 **2,84**	2,04 **2,74**	1,99 **2,66**	1,93 **2,55**	1,89 **2,47**	1,84 **2,38**	1,79 **2,29**	1,76 **2,24**	1,72 **2,16**	1,69 **2,13**	1,66 **2,07**	1,64 **2,03**	1,62 **2,01**
32	4,15 **7,50**	3,30 **5,34**	2,90 **4,46**	2,67 **3,97**	2,51 **3,66**	2,40 **3,42**	2,32 **3,25**	2,25 **3,12**	2,19 **3,01**	2,14 **2,94**	2,10 **2,86**	2,07 **2,80**	2,02 **2,70**	1,97 **2,62**	1,91 **2,51**	1,86 **2,42**	1,82 **2,34**	1,76 **2,25**	1,74 **2,20**	1,69 **2,12**	1,67 **2,08**	1,64 **2,02**	1,61 **1,98**	1,59 **1,96**
34	4,13 **7,44**	3,28 **5,29**	2,88 **4,42**	2,65 **3,93**	2,49 **3,61**	2,38 **3,38**	2,30 **3,21**	2,23 **3,08**	2,17 **2,97**	2,12 **2,89**	2,08 **2,82**	2,05 **2,76**	2,00 **2,66**	1,95 **2,58**	1,89 **2,47**	1,84 **2,38**	1,80 **2,30**	1,74 **2,21**	1,71 **2,15**	1,67 **2,08**	1,64 **2,04**	1,61 **1,98**	1,59 **1,94**	1,57 **1,91**
36	4,11 **7,39**	3,26 **5,25**	2,86 **4,38**	2,63 **3,89**	2,48 **3,58**	2,36 **3,35**	2,28 **3,18**	2,21 **3,04**	2,15 **2,94**	2,10 **2,86**	2,06 **2,78**	2,03 **2,72**	1,89 **2,62**	1,93 **2,54**	1,87 **2,43**	1,82 **2,35**	1,78 **2,26**	1,72 **2,17**	1,69 **2,12**	1,65 **2,04**	1,62 **2,00**	1,59 **1,94**	1,56 **1,90**	1,55 **1,87**
38	4,10 **7,35**	3,25 **5,21**	2,85 **4,34**	2,62 **3,86**	2,46 **3,54**	2,35 **3,32**	2,26 **3,15**	2,19 **3,02**	2,14 **2,91**	2,09 **2,82**	2,05 **2,75**	2,02 **2,69**	1,96 **2,59**	1,92 **2,51**	1,85 **2,40**	1,80 **2,32**	1,76 **2,22**	1,71 **2,14**	1,67 **2,08**	1,63 **2,00**	1,60 **1,97**	1,57 **1,90**	1,54 **1,86**	1,53 **1,84**
40	4,08 **7,31**	3,23 **5,18**	2,84 **4,31**	2,61 **3,83**	2,45 **3,51**	2,34 **3,29**	2,25 **3,12**	2,18 **2,99**	2,12 **2,88**	2,07 **2,80**	2,04 **2,73**	2,00 **2,66**	1,95 **2,56**	1,90 **2,49**	1,84 **2,37**	1,79 **2,29**	1,74 **2,20**	1,69 **2,11**	1,66 **2,05**	1,61 **1,97**	1,59 **1,94**	1,55 **1,88**	1,53 **1,84**	1,51 **1,81**
42	4,07 **7,27**	3,22 **5,15**	2,83 **4,29**	2,59 **3,80**	2,44 **3,49**	2,32 **3,26**	2,24 **3,10**	2,17 **2,96**	2,11 **2,86**	2,06 **2,77**	2,02 **2,70**	1,99 **2,64**	1,94 **2,54**	1,89 **2,46**	1,82 **2,35**	1,78 **2,26**	1,73 **2,17**	1,68 **2,08**	1,64 **2,02**	1,60 **1,94**	1,57 **1,91**	1,54 **1,85**	1,51 **1,80**	1,49 **1,78**
44	4,06 **7,24**	3,21 **5,12**	2,82 **4,26**	2,58 **3,78**	2,43 **3,46**	2,31 **3,24**	2,23 **3,07**	2,16 **2,94**	2,10 **2,84**	2,05 **2,75**	2,01 **2,68**	1,98 **2,62**	1.92 **2,52**	1,88 **2,44**	1,81 **2,32**	1,76 **2,24**	1,72 **2,15**	1,66 **2,06**	1,63 **2,00**	1,58 **1,92**	1,56 **1,88**	1,52 **1,82**	1,50 **1,78**	1,48 **1,75**
46	4,05 **7,21**	3,20 **5,10**	2,81 **4,24**	2,57 **3,76**	2,42 **3,44**	2,30 **3,22**	2,22 **3,05**	2,14 **2,92**	2,09 **2,82**	2,04 **2,73**	2,00 **2,66**	1,97 **2,60**	1,91 **2,50**	1,87 **2,42**	1,80 **2,30**	1,75 **2,22**	1,71 **2,13**	1,65 **2,04**	1,62 **1,98**	1,57 **1,90**	1,54 **1,86**	1,51 **1,80**	1,48 **1,76**	1,46 **1,72**
48	4,04 **7,19**	3,19 **5,08**	2,80 **4,22**	2,56 **3,74**	2,41 **3,42**	2,30 **3,20**	2,21 **3,04**	2,14 **2,90**	2,08 **2,80**	2,03 **2,71**	1,99 **2,64**	1,96 **2,58**	1,90 **2,48**	1,86 **2,40**	1,79 **2,28**	1,74 **2,20**	1,70 **2,11**	1,64 **2,02**	1,61 **1,96**	1,56 **1,88**	1,53 **1,84**	1,50 **1,78**	1,47 **1,73**	1,45 **1,70**
50	4,03 **7,17**	3,18 **5,06**	2,79 **4,20**	2,56 **3,72**	2,40 **3,41**	2,29 **3,18**	2,20 **3,02**	2,13 **2,88**	2,07 **2,78**	2,02 **2,70**	1,98 **2,62**	1,95 **2,56**	1,90 **2,46**	1,85 **2,39**	1,78 **2,26**	1,74 **2,18**	1,69 **2,10**	1,63 **2,00**	1,60 **1,94**	1,55 **1,86**	1,52 **1,82**	1,48 **1,76**	1,46 **1,71**	1,44 **1,68**
55	4,02 **7,12**	3,17 **5,01**	2,78 **4,16**	2,54 **3,68**	2,38 **3,37**	2,27 **3,15**	2,18 **2,98**	2,11 **2,85**	2,05 **2,75**	2,00 **2,66**	1,97 **2,59**	1,93 **2,53**	1,88 **2,43**	1,83 **2,35**	1,76 **2,23**	1,72 **2,15**	1,67 **2,06**	1,61 **1,96**	1,58 **1,90**	1,52 **1,82**	1,50 **1,78**	1,46 **1,71**	1,43 **1,66**	1,41 **1,64**
60	4,00 **7,08**	3,15 **4,98**	2,76 **4,13**	2,52 **3,65**	2,37 **3,34**	2,25 **3,12**	2,17 **2,95**	2,10 **2,82**	2,04 **2,72**	1,99 **2,63**	1,95 **2,56**	1,92 **2,50**	1,86 **2,40**	1,81 **2,32**	1,75 **2,20**	1,70 **2,12**	1,65 **2,03**	1,59 **1,93**	1,56 **1,87**	1,50 **1,79**	1,48 **1,74**	1,44 **1,68**	1,41 **1,63**	*1,39 **1,60**
65	3,99 **7,04**	3,14 **4,95**	2,75 **4,10**	2,51 **3,62**	2,36 **3,31**	2,24 **3,09**	2,15 **2,93**	2,08 **2,79**	2,02 **2,70**	1,98 **2,61**	1,94 **2,54**	1,90 **2,47**	1,85 **2,37**	1,80 **2,30**	1,73 **2,18**	1,68 **2,09**	1,63 **2,00**	1,57 **1,90**	1,54 **1,84**	1,49 **1,76**	1,46 **1,71**	1,42 **1,64**	1,39 **1,60**	1,37 **1,56**
70	3,98 **7,01**	3,13 **4,92**	2,74 **4,08**	2,50 **3,60**	2,35 **3,29**	2,32 **3,07**	2,14 **2,91**	2,07 **2,77**	2,01 **2,67**	1,97 **2,59**	1,93 **2,51**	1,89 **2,45**	1,84 **2,35**	1,79 **2,28**	1,72 **2,15**	1,67 **2,07**	1,62 **1,98**	1,56 **1,88**	1,53 **1,82**	1,47 **1,74**	1,45 **1,69**	1,40 **1,63**	1,37 **1,56**	1,35 **1,53**
80	3,96 **6,96**	3,11 **4,88**	2,72 **4,04**	2,48 **3,56**	2,33 **3,25**	2,21 **3,04**	2,12 **2,87**	2,05 **2,74**	1,99 **2,64**	1,95 **2,55**	1,91 **2,48**	1,88 **2,41**	1,82 **2,32**	1,77 **2,24**	1,70 **2,11**	1,65 **2,03**	1,60 **1,94**	1,54 **1,84**	1,51 **1,78**	1,45 **1,70**	1,42 **1,65**	1,38 **1,57**	1,35 **1,52**	1,32 **1,49**
100	3,94 **6,90**	3,09 **4,82**	2,70 **3,98**	2,46 **3,51**	2,30 **3,20**	2,19 **2,99**	2,10 **2,82**	2,03 **2,69**	1,97 **2,59**	1,92 **2,51**	1,88 **2,43**	1,85 **2,36**	1,79 **2,26**	1,75 **2,19**	1,68 **2,06**	1,63 **1,98**	1,57 **1,89**	1,51 **1,79**	1,48 **1,73**	1,42 **1,64**	1,39 **1,59**	1,34 **1,51**	1,30 **1,46**	1,28 **1,43**
125	3,92 **6,84**	3,07 **4,78**	2,68 **3,94**	2,44 **3,47**	2,29 **3,17**	2,17 **2,95**	2,08 **2,79**	2,01 **2,65**	1,95 **2,56**	1,90 **2,47**	1,86 **2,40**	1,83 **2,33**	1,77 **2,23**	1,72 **2,15**	1,65 **2,03**	1,60 **1,94**	1,55 **1,85**	1,49 **1,75**	1,45 **1,68**	1,39 **1,59**	1,36 **1,54**	1,31 **1,46**	1,27 **1,40**	1,25 **1,37**
150	3,91 **6,81**	3,06 **4,75**	2,67 **3,91**	2,43 **3,44**	2,27 **3,13**	2,16 **2,92**	2,07 **2,76**	2,00 **2,62**	1,94 **2,53**	1,89 **2,44**	1,85 **2,37**	1,82 **2,30**	1,76 **2,20**	1,71 **2,12**	1,64 **2,00**	1,59 **1,91**	1,54 **1,83**	1,47 **1,72**	1,44 **1,66**	1,37 **1,56**	1,34 **1,51**	1,29 **1,43**	1,25 **1,37**	1,22 **1,33**
200	3,89 **6,76**	3,04 **4,71**	2,65 **3,88**	2,41 **3,41**	2,26 **3,11**	2,14 **2,90**	2,05 **2,73**	1,98 **2,60**	1,92 **2,50**	1,87 **2,41**	1,83 **2,34**	1,80 **2,28**	1,74 **1,17**	1,69 **2,09**	1,62 **1,97**	1,57 **1,88**	1,52 **1,79**	1,45 **1,69**	1,42 **1,62**	1,35 **1,53**	1,32 **1,48**	1,26 **1,39**	1,22 **1,33**	1,19 **1,28**
400	3,86 **6,70**	3,02 **4,66**	2,62 **3,83**	2,39 **3,36**	2,23 **3,06**	2,12 **2,85**	2,03 **2,69**	1,96 **2,55**	1,90 **2,46**	1,85 **2,37**	1,81 **2,29**	1,78 **2,23**	1,72 **2,12**	1,67 **2,04**	1,60 **1,92**	1,54 **1,84**	1,49 **1,74**	1,42 **1,64**	1,38 **1,57**	1,32 **1,47**	1,28 **1,42**	1,22 **1,32**	1,16 **1,24**	1,13 **1,19**
1000	3,85 **6,66**	3,00 **4,62**	2,61 **3,80**	2,38 **3,34**	2,22 **3,04**	2,10 **2,82**	2,02 **2,66**	1,95 **2,53**	1,89 **2,43**	1,84 **2,34**	1,80 **2,26**	1,76 **2,20**	1,70 **2,09**	1,65 **2,01**	1,58 **1,89**	1,53 **1,81**	1,47 **1,71**	1,41 **1,61**	1,36 **1,54**	1,30 **1,44**	1,26 **1,38**	1,19 **1,28**	1,13 **1,19**	1,08 **1,11**
	3,84 **6,64**	2,99 **4,60**	2,60 **3,78**	2,37 **3,32**	2,21 **3,02**	2,09 **2,80**	2,01 **2,64**	1,94 **2,51**	1,88 **2,41**	1,83 **2,32**	1,79 **2,24**	1,75 **2,18**	1,69 **2,07**	1,64 **1,99**	1,57 **1,87**	1,52 **1,79**	1,46 **1,69**	1,40 **1,59**	1,35 **1,52**	1,28 **1,41**	1,24 **1,36**	1,17 **1,25**	1,11 **1,15**	1,00 **1,00**

TABELA X — Coeficiente de correlação de postos.

	Nível de significância (teste unilateral)	
n	0,05	0,01
4	1,000	
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,746
12	0,504	0,701
14	0,456	0,645
16	0,425	0,601
18	0,399	0,564
20	0,377	0,534
22	0,359	0,508
24	0,343	0,485
26	0,329	0,465
28	0,317	0,448
30	0,306	0,432

TABELA XI — Dígitos aleatórios.

31	75	15	72	60	68	98	00	53	39	15	47	04	83	55	88	65	12	25	96	03	15	21	91	21
88	49	29	93	82	14	45	40	45	04	20	09	49	89	77	74	84	39	34	13	22	10	97	85	08
30	93	44	77	44	07	48	18	38	28	73	78	80	65	33	28	59	72	04	05	94	20	52	03	80
22	88	84	88	93	27	49	99	87	48	60	53	04	51	28	74	02	28	46	17	82	03	71	02	68
78	21	21	69	93	35	90	29	13	86	44	37	21	54	86	65	74	11	40	14	87	48	13	72	20
41	84	98	45	47	46	85	05	23	26	34	67	75	83	00	74	91	06	43	45	19	32	58	15	49
46	35	23	30	49	69	24	89	34	60	45	30	50	75	21	61	31	83	18	55	14	41	37	09	51
11	08	79	62	94	14	01	33	17	92	59	74	76	72	77	76	50	33	45	13	39	66	37	75	44
52	70	10	83	37	56	30	38	73	15	16	52	06	96	76	11	65	49	98	93	02	18	16	81	61
57	27	53	68	98	81	30	44	85	85	68	65	22	73	76	92	85	25	58	66	88	44	80	35	84
20	85	77	31	56	70	28	42	43	26	79	37	59	52	20	01	15	96	32	67	10	62	24	83	91
15	63	38	49	24	90	41	59	36	14	33	52	12	66	65	55	82	34	76	41	86	22	53	17	04
92	69	44	82	97	39	90	40	21	15	59	58	94	90	67	66	82	14	15	75	49	76	70	40	37
77	61	31	90	19	88	15	20	00	80	20	55	49	14	09	96	27	74	82	57	50	81	69	76	16
38	68	83	24	86	45	13	46	35	45	59	40	47	20	59	43	94	75	16	80	43	85	25	96	93
25	16	30	18	89	70	01	41	50	21	41	29	06	73	12	71	85	71	59	57	68	97	11	14	03
65	25	10	76	29	37	23	93	32	95	05	87	00	11	19	92	78	42	63	40	18	47	76	56	22
36	81	54	36	25	18	63	73	75	09	82	44	49	90	05	04	92	17	37	01	14	70	79	39	97
64	39	71	16	92	05	32	78	21	62	20	24	78	17	59	45	19	72	53	32	83	74	52	25	67
04	51	52	56	24	95	09	66	79	46	48	46	08	55	58	15	19	11	87	82	16	93	03	33	61
15	88	09	22	61	17	29	28	81	90	61	78	14	88	98	92	52	52	12	83	88	58	16	00	98
71	92	60	08	19	59	14	40	02	24	30	57	09	01	94	18	32	90	69	99	26	85	71	92	38
64	42	52	81	08	16	55	41	60	16	00	04	28	32	29	10	33	33	61	68	65	61	79	48	34
79	78	22	39	24	49	44	03	04	32	81	07	73	15	43	95	21	66	48	65	13	65	85	10	81
36	33	77	45	38	44	55	36	46	72	90	96	04	18	49	93	86	54	46	08	93	17	63	48	51
05	24	92	93	29	19	71	59	40	82	14	73	88	66	67	43	70	86	63	54	93	69	22	55	27
56	46	39	93	80	38	79	38	57	74	19	05	61	39	39	46	06	22	76	47	66	14	66	32	10
96	29	63	31	21	54	19	63	41	08	75	81	48	59	86	71	17	11	51	02	28	99	26	31	65
98	38	03	62	69	60	01	40	72	01	62	44	84	63	85	42	17	58	83	50	46	18	24	91	26
52	56	76	43	50	16	31	55	39	69	80	39	58	11	14	54	35	86	45	78	47	26	91	57	47
78	45	89	08	30	25	95	59	92	36	43	28	69	10	64	99	96	99	51	44	64	42	47	73	77
49	55	32	42	41	08	15	08	95	35	08	70	39	10	41	77	32	38	10	79	45	12	79	36	86
32	15	10	70	75	83	15	51	02	52	73	10	08	86	18	23	89	18	74	18	45	41	72	02	68
11	31	45	03	63	26	86	02	77	99	49	41	68	35	34	19	18	70	80	59	76	67	70	21	10
12	36	47	12	10	87	05	25	02	41	90	78	59	78	89	81	39	95	81	30	64	43	90	56	14
09	18	82	00	97	32	82	53	95	27	04	22	08	63	04	83	38	98	73	74	64	27	85	80	44
90	04	58	54	97	51	98	15	06	54	94	93	88	19	97	91	87	07	61	50	68	47	66	46	59
73	18	95	02	07	47	67	72	62	69	62	29	06	44	64	27	12	46	70	18	41	36	18	27	60
75	76	87	64	90	20	97	18	17	49	90	42	91	22	72	95	37	50	58	71	93	82	34	31	78
54	01	64	40	56	66	28	13	10	03	00	68	22	73	98	20	71	45	32	95	07	70	61	78	13
08	35	86	99	10	78	54	24	27	85	13	66	15	88	73	04	61	89	75	53	31	22	30	84	20
28	30	60	32	64	81	33	31	05	91	40	51	00	78	93	32	60	46	04	75	94	11	90	18	40
53	84	08	62	33	81	59	41	36	28	51	21	59	02	90	28	46	66	87	95	77	76	22	07	91
91	75	75	37	41	61	61	36	22	69	50	26	39	02	12	55	78	17	65	14	83	48	34	70	55
89	41	59	26	94	00	39	75	83	91	12	60	71	76	46	48	94	97	23	06	94	54	13	74	08
77	51	30	38	20	86	83	42	99	01	68	41	48	27	74	51	90	81	39	80	72	89	35	55	07
19	50	23	71	74	69	97	92	02	88	55	21	02	97	73	74	28	77	52	51	65	34	46	74	15
21	81	85	93	13	93	27	88	17	57	05	68	67	31	56	07	08	28	50	46	31	85	33	84	52
51	47	46	64	99	68	10	72	36	21	94	04	99	13	45	42	83	60	91	91	08	00	74	54	49
99	55	96	83	31	62	53	52	41	70	69	77	71	28	30	74	81	97	81	42	43	86	07	28	34
60	31	14	28	24	37	30	14	26	78	45	99	04	32	42	17	37	45	20	03	70	70	77	02	14
49	73	97	14	84	92	00	39	80	86	76	66	87	32	09	59	20	21	19	73	02	90	23	32	50
78	62	65	15	94	16	45	39	46	14	39	01	49	70	66	83	01	20	98	32	25	57	17	76	28
66	69	21	39	86	99	83	70	05	82	81	23	24	49	87	09	50	49	64	12	90	19	37	95	68
44	07	12	80	91	07	36	29	77	03	76	44	74	25	37	98	52	49	78	31	65	70	40	95	14
41	46	88	51	49	49	55	41	79	94	14	92	43	96	50	95	29	40	05	56	70	48	10	69	05
94	55	93	75	59	49	67	85	31	19	70	31	20	56	82	66	98	63	40	99	74	47	42	07	40
41	61	57	03	60	64	11	45	86	60	90	85	06	46	18	80	62	05	17	90	11	43	63	80	72
50	27	39	31	13	41	79	48	68	61	24	78	18	96	83	55	41	18	56	67	77	53	59	98	92
41	39	68	05	04	90	67	00	82	89	40	90	20	50	69	95	08	30	67	83	28	10	25	78	16

TABELA XI — (continuação)

25	80	72	42	60	71	52	97	89	20	72	68	20	73	85	90	72	65	71	66	98	88	40	85	83
06	17	09	79	65	88	30	29	80	41	21	44	34	18	08	68	98	48	36	20	89	74	79	88	82
60	80	85	44	44	74	41	28	11	05	01	17	62	88	38	36	42	11	64	89	18	05	95	10	61
80	94	04	48	93	10	40	83	62	22	80	58	27	19	44	92	63	84	03	33	67	05	41	60	67
19	51	69	01	20	46	75	97	16	43	13	17	75	52	92	21	03	68	28	08	77	50	19	74	27
49	38	65	44	80	23	60	42	35	54	21	78	54	11	01	91	17	81	01	74	29	42	09	04	38
06	31	28	89	40	15	99	56	93	21	47	45	86	48	09	98	18	98	18	51	29	65	18	42	15
60	94	20	03	07	11	89	79	26	74	40	40	56	80	32	96	71	75	42	44	10	70	14	13	93
92	32	99	89	32	78	28	44	63	47	71	20	99	20	61	39	44	89	31	36	25	72	20	85	64
77	93	66	35	74	31	38	45	19	24	85	56	12	96	71	58	13	71	78	20	22	75	13	65	18
91	30	70	69	91	19	07	22	42	10	36	69	95	37	28	28	82	53	57	93	28	97	66	62	52
68	43	49	46	88	84	47	31	36	22	62	12	69	84	08	12	84	38	25	90	09	81	59	31	46
48	90	81	58	77	54	74	52	45	91	35	70	00	47	54	83	82	45	26	92	54	13	05	51	60
06	91	34	51	97	42	67	27	86	01	11	88	30	95	28	63	01	19	89	01	14	97	44	03	44
10	45	51	60	19	14	21	03	37	12	91	34	23	78	21	88	32	58	08	51	43	66	77	08	83
12	88	39	73	43	65	02	76	11	84	04	28	50	13	92	17	97	41	50	77	90	71	22	67	69
21	77	83	09	76	38	80	73	69	61	31	64	94	20	96	63	28	10	20	23	08	81	64	74	49
19	52	35	95	15	65	12	25	96	59	86	28	36	82	58	69	57	21	37	98	16	43	59	15	29
67	24	55	26	70	35	58	31	65	63	79	24	68	66	86	76	46	33	42	22	26	65	59	08	02
60	58	44	73	77	07	50	03	79	92	45	13	42	65	29	26	76	08	36	37	41	32	64	43	44
53	85	34	13	77	36	06	69	48	50	58	83	87	38	59	49	36	47	33	31	96	24	04	36	42
24	63	73	87	36	74	38	48	93	42	52	62	30	79	92	12	36	91	86	01	03	74	28	38	73
83	08	01	24	51	38	99	22	28	15	07	75	95	17	77	97	37	72	75	85	51	97	23	78	67
16	44	42	43	34	36	15	19	90	73	27	49	37	09	39	85	13	03	25	52	54	84	65	47	59
60	79	01	81	57	57	17	86	57	62	11	16	17	85	76	45	81	95	29	79	65	13	00	48	60
94	01	54	68	74	32	44	44	82	77	59	82	09	61	63	64	65	42	58	43	41	14	54	28	20
74	10	88	82	22	88	57	07	40	15	25	70	49	10	35	01	75	51	47	50	48	96	83	86	03
62	88	08	78	73	95	16	05	92	21	22	30	49	03	14	72	87	71	73	34	39	28	30	41	49
11	74	81	21	02	80	58	04	18	67	17	71	05	96	21	06	55	40	78	50	73	95	07	95	52
17	94	40	56	00	60	47	80	33	43	25	85	25	89	05	57	21	63	96	18	49	85	69	93	26
66	06	74	27	92	95	04	35	26	80	46	78	05	64	87	09	97	15	94	81	37	00	62	21	86
54	24	49	10	30	45	54	77	08	18	59	84	99	61	69	61	45	92	16	47	87	41	71	71	98
30	94	55	75	89	31	73	25	72	60	47	67	00	76	54	46	37	62	53	66	94	74	64	95	80
69	17	03	74	03	86	99	59	03	07	94	30	47	18	03	26	82	50	55	11	12	45	99	13	14
08	34	58	89	75	35	84	18	57	71	08	10	55	99	87	87	11	22	14	76	14	71	37	11	81
27	76	74	35	84	85	30	18	89	77	29	49	06	97	14	73	03	54	12	07	74	69	90	93	10
13	02	51	43	38	54	06	61	52	43	47	72	46	67	33	47	43	14	39	05	31	04	85	66	99
80	21	73	62	92	98	52	52	43	35	24	43	22	48	96	43	27	75	88	74	11	46	61	60	82
10	87	56	20	04	90	39	16	11	05	57	41	10	63	68	53	85	63	07	43	08	67	08	47	41
54	12	75	73	26	26	62	91	90	87	24	47	28	87	79	30	54	02	78	86	61	73	27	54	54
33	71	34	80	07	93	58	47	28	69	51	92	66	47	21	58	30	32	98	22	93	17	49	39	72
85	27	48	68	93	11	30	32	92	70	28	83	43	41	37	73	51	59	04	00	71	14	84	36	43
84	13	38	96	40	44	03	55	21	66	73	85	27	00	91	61	22	26	05	61	62	32	71	84	23
56	73	21	62	34	17	39	59	61	31	10	12	39	16	22	85	49	65	75	60	81	60	41	88	80
65	13	85	68	06	87	64	88	52	61	34	31	36	58	61	45	87	52	10	69	85	64	44	72	77
38	00	10	21	76	81	71	91	17	11	71	60	29	29	37	74	21	96	40	49	65	58	44	96	98
37	40	29	63	97	01	30	47	75	86	56	27	11	00	86	47	32	46	26	05	40	03	03	74	38
97	12	54	03	48	87	08	33	14	17	21	81	53	92	50	75	23	76	20	47	15	50	12	95	78
21	82	64	11	34	47	14	33	40	72	64	63	88	59	02	49	13	90	64	41	03	85	65	45	52
73	13	54	27	42	95	71	90	90	35	85	79	47	42	96	08	78	98	81	56	64	69	11	92	02
07	63	87	79	29	03	06	11	80	72	96	20	74	41	56	23	82	19	95	38	04	71	36	69	94
60	52	88	34	41	07	95	41	98	14	59	17	52	06	95	05	53	35	21	39	61	21	20	64	55
83	59	63	56	55	06	95	89	29	83	05	12	80	97	19	77	43	35	37	83	92	30	15	04	98
10	85	06	27	46	99	59	91	05	07	13	49	90	63	19	53	07	57	18	39	06	41	01	93	62
39	82	09	89	52	43	62	26	31	47	64	42	18	08	14	43	80	00	93	51	31	02	47	31	67
59	58	00	64	78	75	56	97	88	00	88	83	55	44	86	23	76	80	61	56	04	11	10	84	08
38	50	80	73	41	23	79	34	87	63	90	82	29	70	22	17	71	90	42	07	95	95	44	99	53
30	69	27	06	68	94	68	81	61	27	56	19	68	00	91	82	06	76	34	00	05	46	26	92	00
65	44	39	56	59	18	28	82	74	37	49	63	22	40	41	08	33	76	56	76	96	29	99	08	36
27	26	75	02	64	13	19	27	22	94	07	47	74	46	06	17	98	54	89	11	97	34	13	03	58

TABELA XI — (continuação)

38	10	17	77	56	11	65	71	38	97	95	88	95	70	67	47	64	81	38	85	70	66	99	34	06
39	64	16	94	57	91	33	92	25	02	92	61	38	97	19	11	94	75	62	03	19	32	42	05	04
84	05	44	04	55	99	39	66	36	80	67	66	76	06	31	69	18	19	68	45	38	52	51	16	00
47	46	80	35	77	57	64	96	32	66	24	70	07	15	94	14	00	42	31	53	69	24	90	57	47
43	32	13	13	70	28	97	72	38	96	76	47	96	85	62	62	34	20	75	89	08	89	90	59	85
64	28	16	18	26	18	55	56	49	37	13	17	33	33	65	78	85	11	64	99	87	06	41	30	75
66	84	77	04	95	32	35	00	29	85	86	71	63	87	46	26	31	37	74	63	55	38	77	26	81
72	46	13	32	30	21	52	95	34	24	92	58	10	22	62	78	43	86	62	76	18	39	67	35	38
21	03	29	10	50	13	05	81	62	18	12	47	05	65	00	15	29	27	61	39	59	52	65	21	13
95	36	26	70	11	06	65	11	61	36	01	01	60	08	57	55	01	85	63	74	35	82	47	17	08
40	71	29	73	80	10	40	45	54	52	34	03	06	07	26	75	21	11	02	71	36	63	36	84	24
58	27	56	17	64	97	58	65	47	16	50	25	94	63	45	87	19	54	60	92	26	78	76	09	39
89	51	41	17	88	68	22	42	34	17	73	95	97	61	45	30	34	24	02	77	11	04	97	20	49
15	47	25	06	69	48	13	93	67	32	46	87	43	70	88	73	46	50	98	19	58	86	93	52	20
12	12	08	61	24	51	24	74	43	02	60	88	35	21	09	21	43	73	67	86	49	22	67	78	37
03	99	11	04	61	93	71	61	68	94	66	08	32	46	53	84	60	95	82	32	88	61	81	91	61
38	55	59	55	54	32	88	65	97	80	08	35	56	08	60	29	73	54	77	62	71	29	92	38	53
17	54	67	37	04	92	05	24	62	15	55	12	12	92	81	59	07	60	79	36	27	95	45	89	09
32	64	35	28	61	95	81	90	68	31	00	91	19	89	36	76	35	59	37	79	80	86	30	05	14
69	57	26	87	77	39	51	03	59	05	14	06	04	06	19	29	54	96	96	16	33	56	46	07	80
24	12	26	65	91	27	69	90	64	94	14	84	54	66	72	61	95	87	71	00	90	89	97	57	54
61	19	63	02	31	92	96	26	17	73	41	83	95	53	82	17	26	77	09	43	78	03	87	02	67
30	53	22	17	04	10	27	41	22	02	39	68	52	33	09	10	06	16	88	29	55	98	66	64	85
03	78	89	75	99	75	86	72	07	17	74	41	65	31	66	35	20	83	33	74	87	53	90	88	23
48	22	86	33	79	85	78	34	76	19	53	15	26	74	33	35	66	35	29	72	16	81	86	03	11
60	36	59	46	53	35	07	53	39	49	42	61	42	92	97	01	91	82	83	16	98	95	37	32	31
83	79	94	24	02	56	62	33	44	42	34	99	44	13	74	70	07	11	47	36	09	95	81	80	65
32	96	00	74	05	36	40	98	32	32	99	38	54	16	00	11	13	30	75	86	15	91	70	62	53
19	32	25	38	45	57	62	05	26	06	66	49	76	86	46	78	13	86	65	59	19	64	09	94	13
11	22	09	47	47	07	39	93	74	08	48	50	92	39	29	27	48	24	54	76	85	24	43	51	59
21	44	58	27	93	24	83	19	32	41	14	19	97	62	68	70	88	36	80	02	03	82	91	74	43
72	51	37	64	00	52	22	59	23	48	62	30	89	84	81	29	74	43	31	65	33	14	16	10	20
71	47	94	50	27	76	16	05	74	11	13	78	01	36	32	52	30	87	77	62	88	87	43	36	97
83	21	05	14	66	09	08	85	03	95	26	74	30	53	06	21	70	67	00	01	99	43	98	07	67
68	74	99	51	48	94	89	77	86	36	96	75	00	90	24	94	53	89	11	43	96	69	36	18	86
05	18	47	57	63	47	07	58	81	58	05	31	35	34	39	14	90	80	88	30	60	09	62	15	51
13	65	16	25	46	96	89	22	52	40	47	51	15	84	83	87	34	27	88	18	07	85	53	92	69
00	56	62	12	20	00	29	22	40	69	25	07	22	95	19	52	54	85	40	91	21	28	22	12	96
50	95	81	76	95	58	07	26	89	90	60	32	99	59	55	71	58	66	34	17	35	94	76	78	07
57	62	16	45	47	46	85	03	79	81	38	52	70	90	37	64	75	60	33	24	04	98	68	36	66
09	28	22	58	44	79	13	97	84	35	35	42	84	35	61	69	79	96	33	14	12	99	19	35	16
23	39	49	42	06	93	43	23	78	36	94	91	92	68	46	02	55	57	44	10	94	91	54	81	99
05	28	03	74	70	93	62	20	43	45	15	09	21	95	10	18	09	41	66	13	78	23	45	00	01
95	49	19	79	76	38	30	63	21	92	82	63	95	46	24	72	43	49	26	06	23	19	17	46	93
78	52	10	01	04	18	24	87	55	83	90	32	65	07	85	54	03	46	62	51	35	77	41	46	92
96	34	54	45	79	85	93	24	40	53	75	70	42	08	40	86	58	38	39	44	52	45	67	37	66
77	96	33	11	51	32	36	49	16	91	47	35	74	03	38	23	43	52	40	65	08	45	89	53	66
07	52	01	12	94	23	23	80	17	48	41	69	06	73	28	54	81	43	77	77	10	05	74	23	32
38	42	30	23	09	70	70	38	57	36	46	14	81	42	58	29	23	61	21	52	05	08	86	58	25
02	46	36	55	33	21	19	96	05	55	33	92	80	18	17	07	39	68	92	15	30	72	22	21	02
83	76	16	08	73	43	25	38	41	45	60	83	32	59	83	01	29	14	13	49	20	36	80	71	26
14	38	70	63	45	80	85	40	92	79	43	52	90	63	18	38	38	47	47	61	41	19	63	74	80
51	32	19	22	46	80	08	87	70	74	88	72	25	67	36	66	16	44	94	31	66	91	93	16	78
72	47	20	00	08	80	89	01	80	02	94	81	33	19	00	54	15	58	34	36	35	35	25	41	31
05	46	65	53	06	93	12	81	84	64	74	45	79	05	61	72	84	81	18	34	79	98	26	84	16
39	52	87	24	84	82	47	42	55	93	48	54	53	52	47	18	61	91	36	74	18	61	11	92	41
81	61	61	87	11	53	34	24	42	76	75	12	21	17	24	74	62	77	37	07	58	31	91	59	97
07	58	61	61	20	82	64	12	28	20	92	90	41	31	41	32	39	21	97	63	61	19	96	79	40
90	76	70	42	35	13	57	41	72	00	69	90	26	37	42	78	46	42	25	01	18	62	79	08	72
40	18	82	81	93	29	59	38	86	27	94	97	21	15	98	62	09	53	67	87	00	44	15	89	97

TABELA XI — (continuação)

34	41	48	21	57	86	88	75	50	87	19	15	20	00	23	12	30	28	07	83	32	62	46	86	91
63	43	97	53	63	44	98	91	68	22	36	02	40	08	67	76	37	84	16	05	65	96	17	34	88
67	04	90	90	70	93	39	94	55	47	94	45	87	42	84	05	04	14	98	07	20	28	83	40	60
79	49	50	41	46	52	16	29	02	86	54	15	83	42	43	46	97	83	54	82	59	36	29	59	38
91	70	43	05	52	04	73	72	10	31	75	05	19	30	29	47	66	56	43	82	99	78	29	34	78
19	61	27	84	30	11	66	19	47	70	77	60	36	56	69	86	86	81	26	65	30	01	27	59	89
39	14	17	74	00	28	00	06	42	38	73	25	87	17	94	31	34	02	62	56	66	45	33	70	16
64	75	68	04	57	08	74	71	28	36	03	46	95	06	78	03	27	44	34	23	66	67	78	25	56
92	90	15	18	78	56	44	12	29	98	29	71	83	84	47	06	45	32	53	11	07	56	55	37	71
03	55	19	00	70	09	48	39	40	50	45	93	81	81	35	36	90	84	33	21	11	07	35	18	03
98	88	46	62	09	06	83	05	36	56	14	66	35	63	46	71	43	00	49	09	19	81	80	57	07
27	36	98	68	82	53	47	30	76	41	53	63	37	08	63	03	74	81	28	22	19	36	04	90	88
59	06	67	59	74	63	33	52	04	83	43	51	43	74	81	58	27	82	69	67	49	43	54	39	51
91	64	79	37	83	64	16	94	90	22	98	58	80	94	95	49	82	95	90	68	38	83	10	48	38
83	60	59	24	19	39	54	20	77	72	71	56	87	56	73	35	18	58	97	59	44	90	17	42	91
24	89	58	85	30	70	77	43	54	39	46	75	87	04	72	70	20	79	26	75	91	62	36	12	75
35	72	02	65	56	95	59	62	00	94	73	75	08	57	88	34	26	40	17	03	46	83	36	52	48
14	14	15	34	10	38	64	90	63	43	57	25	66	13	42	72	70	97	53	18	90	37	93	75	62
27	41	67	56	70	92	17	67	25	35	93	11	95	60	77	06	88	61	82	44	92	34	43	13	74
82	07	10	74	29	81	00	74	77	49	40	74	45	69	74	23	33	68	88	21	53	84	11	05	36

TABELA XII — Dados de avaliação do professor.

Os dois conjuntos seguintes de dados foram obtidos de um estudo feito na Universidade Estadual da Califórnia, Los Angeles, sob a orientação do professor Mildred Massey. Trata-se da relação entre a avaliação do professor feita por seu aluno e o sucesso deste mesmo aluno no curso. O estudo considerou várias outras variáveis além das que aparecem aqui e levou em consideração dados adicionais de outros cursos. No experimento entraram vários cursos de Administração, Economia e Estatística com vários instrutores em cada curso. Foram aplicados exames comuns a todas as seções do mesmo curso e compararam-se os resultados de cursos diversos.

Estou muitíssimo agradecido ao professor Massey por me haver permitido usar estes dados de seu estudo. Para preservar a confidencialidade do estudo, os dados foram marcados como indicamos abaixo.

As variáveis selecionadas são as seguintes:

$\mathbf{x}_1$ = nota total dos testes; $\mathbf{x}_2$ = nota do exame final;

$\mathbf{x}_3$ = nota total do curso ($\mathbf{x}_1 + \mathbf{x}_2$); $\mathbf{x}_4$ = média;

$\mathbf{x}_5$ = nota da avaliação do professor (vai de 1 a 5).

Curso A

x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
57	62	119	2,6	5	129	53	182	2,0	4	174	236	410	3,8	4
144	172	316	2,2	3	113	100	213	2,2	3	91	110	201	2,5	5
110	178	288	2,5	3	121	116	237	2,6	3	147	181	328	2,3	3
159	224	383	3,5	4	121	120	241	2,4	3	166	225	391	2,9	5
113	92	205	2,0	5	121	219	340	3,0	5	110	131	241	3,4	5
140	172	312	2,8	5	110	72	182	2,7	3	98	100	198	3,0	4
76	110	186	2,3	4	151	173	324	2,4	3	98	100	198	2,0	5
121	124	245	2,5	4	98	72	170	2,6	3	124	117	241	2,2	3
159	200	359	2,0	5	140	208	348	3,6	3	132	176	308	3,0	4
129	140	269	2,0	4	151	181	332	3,5	2	140	133	273	2,0	3
170	221	391	3,3	5	110	147	257	2,0	4	91	75	166	2,0	2
125	140	265	3,3	4	117	207	324	3,5	3	151	118	269	2,5	4
136	156	292	3,2	4	140	140	280	2,0	3	117	152	269	2,1	5
106	107	213	2,3	4	102	111	213	2,7	3	144	121	265	1,5	4
121	69	190	2.0	4	140	164	304	3,4	3	87	107	194	1,9	3
109	148	257	2,4	3	98	68	166	2,9	1	155	141	296	2,9	4
147	126	273	2,4	4	147	189	336	2,7	1	144	164	308	2,0	4
144	113	257	2,5	4	144	208	352	2,9	3	140	152	292	2,9	5
144	156	300	2,0	4	136	125	261	1,6	1	117	140	257	2,4	4
49	117	166	2,6	4	110	182	292	2,6	5	121	128	249	2,9	2
151	189	340	3,3	2	174	225	399	3,5	1	163	196	359	3,0	3
110	182	292	2,7	4	83	128	211	2,7	1	159	216	375	3,6	3
144	231	375	4,0	4	136	144	280	3,1	2					
98	226	324	2,5	3	121	120	241	2,1	3					
91	162	253	2,0	3	155	208	363	2,6	3					
132	184	316	2,8	5	129	163	292	2,9	4					
136	89	225	2,7	5	113	203	316	3,0	4					
87	55	142	2,2	4	163	181	344	2,8	5					
159	161	320	3,2	3	147	181	328	2,0	4					
140	172	312	3,5	4	147	141	288	2,8	1					

TABELA XII — (continuação)

Curso B

x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
97	148	245	2,0	3	128	161	289	2,5	4
170	169	339	3,5	4	131	180	311	3,0	4
100	170	270	2,3	5	111	117	228	3,0	3
75	100	175	1,9	4	86	97	183	2,5	4
75	106	181	2,5	4	145	191	336	3,7	5
100	83	183	2,5	1	128	144	272	3,4	5
81	114	195	2,2	3	136	125	261	2,4	4
119	112	231	3,0	3	131	153	284	2,9	4
145	161	306	2,5	3	114	139	253	2,2	3
114	150	264	3,6	1	150	153	303	3,0	3
136	159	295	2,0	2	83	92	175	3,0	3
72	95	167	2,8	3	108	89	197	2,9	2
103	142	245	2,0	3	83	112	195	2,3	3
120	141	261	3,0	3	133	120	253	3,1	3
106	94	200	2,0	3	106	133	239	2,9	3
86	120	206	3,0	5	117	103	220	2,5	5
86	95	181	2,1	3	92	108	200	2,0	4
147	155	302	2,6	5	108	167	275	2,7	3
175	203	378	4,0	3	89	136	225	3,8	3
186	211	397	3,5	5	147	175	322	2,2	4
128	169	297	3,0	4	122	145	267	2,5	3
86	109	195	2,8	3	111	136	247	2,3	5
150	131	281	2,7	3	70	100	170	2,5	5
147	153	300	2,7	3	81	100	181	2,0	5
95	127	222	2,0	5	153	169	322	2,2	5
117	136	253	2,3	1	108	114	222	3,0	3
133	167	300	2,6	4	103	128	231	2,8	4
117	144	261	2,4	3	125	139	264	3,0	4
147	175	322	2,0	5	114	106	220	3,0	3
133	159	292	2,6	2	103	125	228	3,0	4
					111	145	256	3,0	5
					131	164	295	2,0	5
					114	111	225	3,0	5
					117	139	256	2,4	5
					103	153	256	2,8	4

MANUAL DE SOLUÇÕES

Respostas numéricas dependem, freqüentemente, da ordem das operações e da extensão do arredondamento; por isso suas respostas podem diferir um pouco das apresentadas aqui.

CAPÍTULO 2

4. Os pontos de classe são: 156, 159 etc..

5. Os limites de classe são: 0, 1, 2, 3, 4, 5, 6.

7. Inclinada para a direita, porque os estudantes com média menor que C permanecerão fora da escola e, provavelmente, a média não será muito mais alta que o ponto médio entre C e B.

8. Calcule 4 ou 5.

9. $\bar{x} = 4{,}5$.

10. $\bar{x} = 175{,}3$.

11. $\bar{x} = 45{,}4$; $s = 5{,}0$.

12. $\bar{x} = 2{,}09$; $s = 0{,}71$.

13. $s = 2{,}1$.

14. $s = 6{,}6$.

15. (2,4, 6,6) compreende $0{,}1(7) + 12 + 16 + 10 + 4 + 0{,}1\,(5) = 43{,}2$ observações; donde 72%. (0,3, 8,7) exclui $0 + 0{,}8(1) + 1 + 1 = 2{,}8$ observações; por isso 95% caem dentro.

16. (168,7, 181,9) compreende $\frac{0{,}8}{3}$ (90) + 145 + 188 + 180 + 130 + $\frac{0{,}4}{3}$ (90) = 679 observações; donde 68%. (162,1, 188,5) exclui

$4 + 10 + \frac{1,6}{3}(25) + \frac{2}{3}(26) + 4 + 1 + 1 \cong 50,7$ observações; portanto, 95% caem dentro.

19. Estime $\sigma \cong 1$.

21. (a) A média aumentará de 10 pontos mas o desvio-padrão não será afetado.

(b) Tanto a média como o desvio-padrão duplicarão de valor.

23. $\bar{x} = 0$, $s = 2,01$; portanto $(\bar{x} \pm s) = (0 \pm 2,01)$ abarca 100% das observações. Este tipo de distribuição muito afastado do normal tem um desvio-padrão muito grande.

25. Mediana = 45, domínio = 16

26. Mediana = 1,95, domínio = 2,5.

27. Primeiro quartil = 1,6, terceiro quartil = 2,4; por isso o espaço interquartílico = 0,8.

28. (a)

x	95	100	105	110	115	120	125	130	135	140	145	150	155	160	165	170
f	2	2	1	6	8	7	8	2	5	1	1	2	3	1	0	1

(c) Estime 125, (d) $\bar{x} = 124,6$, (e) $\bar{x} = 124,6$, (f) Estime 15-20, (g) $s = 16,9$, (h) $s = 17,0$, (i) Estime 113, 122, 135.

29. (a)

x	8,55	10,55	12,55	14,55	16,55	18,55	20,55	22,55	24,55
f	6	4	10	6	4	2	2	1	2

26,55	28,55	30,55	32,55
1	1	2	1

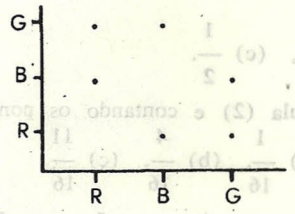
(c) 16, (d) $\bar{x} = 15,9$, (e) $\bar{x} = 16,0$, (f) 6 ou 7, (g) $s = 6,7$, (h) $s = 6,6$, (i) Estime 12, 14, 19.

CAPÍTULO 3

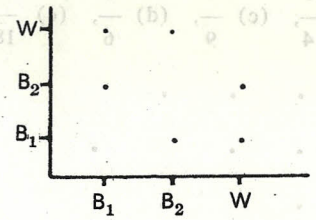
1. Para cada um dos 8 sucessos de 3 lançamentos atribua uma quarta letra Ca ou C para obter as 16 ocorrências possíveis:

CaCaCaCa, CaCaCaC, CaCaCCa, CaCaCC, CaCCaCa, CaCCaC, CCaCaCa, CCaCaC, CaCCCa, CaCCC, CCaCCa, CCaCC, CCCaCa CCCaC, CCCCa, CCCC.

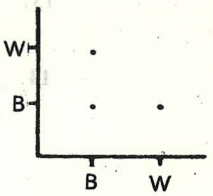
2.



3. (a)



(b).

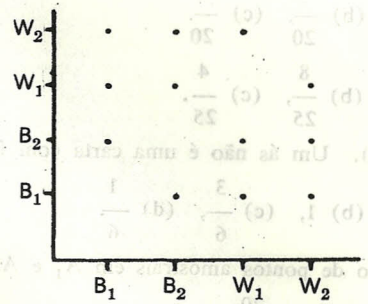


4. $\frac{1}{16}$.

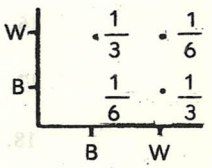
5. $\frac{1}{6}$ e $\frac{1}{9}$.

6. (a) $\frac{1}{16}$, (b) $\frac{1}{3}$.

7.



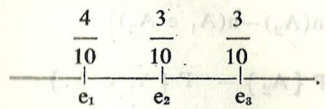
ou



$\frac{1}{12}$ de cada um aqui.

8. Sim, a não ser que houvesse uma subida gradual constante ou uma queda do mercado. O mercado poderia crescer rapidamente por meio de grandes saltos, ainda que a freqüência das altas pudesse ser menor que a freqüência das baixas.

9. (a)



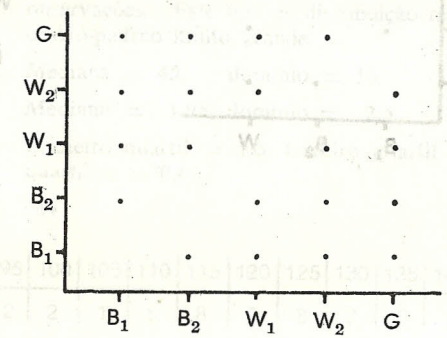
10. (a) $\frac{4}{10}$, (b) $\frac{7}{10}$.

11. (a) $\frac{1}{12}$, (b) $\frac{1}{12}$, (c) $\frac{1}{2}$.

12. Usando a fórmula (2) e contando os pontos amostrais em jogo, as soluções são: (a) $\frac{1}{16}$, (b) $\frac{4}{16}$, (c) $\frac{11}{16}$.

13. (a) $\frac{8}{9}$, (b) $\frac{1}{4}$, (c) $\frac{4}{9}$, (d) $\frac{5}{6}$, (e) $\frac{5}{18}$.

14.



(a) $\frac{2}{20}$, (b) $\frac{8}{20}$, (c) $\frac{4}{20}$.

15. (a) $\frac{4}{25}$, (b) $\frac{8}{25}$, (c) $\frac{4}{25}$.

16. (b) e (d). Um ás não é uma carta com figura.

17. (a) $\frac{4}{6}$, (b) 1, (c) $\frac{3}{6}$, (d) $\frac{1}{6}$.

18. O número de pontos amostrais em A_1 e A_2 é 20; portanto

$$P\{A_1 \text{ ou } A_2\} = \frac{20}{36}.$$

19. O número total de pontos em A_1 e A_2 é dado por $n(A_1) + n(A_2) - n(A_1 \text{ e } A_2)$ porque $n(A_1)+n(A_2)$ conta o número de pontos na interseção de A_1 e A_2 duas vezes; portanto,

$$P\{A_1 \text{ ou } A_2\} = \frac{1}{n}[n(A_1)+n(A_2)-n(A_1 \text{ e } A_2)]$$
$$= P\{A_1\} + P\{A_2\} - P\{A_1 \text{ e } A_2\}.$$

20. $\frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}$

21. (a) $\frac{5}{10} \cdot \frac{4}{9} = \frac{2}{9}$, (b) $\frac{5}{10} \cdot \frac{4}{9} + \frac{3}{10} \cdot \frac{2}{9} + \frac{2}{10} \cdot \frac{1}{9} = \frac{14}{45}$.

22. $P\{2 \text{ caras}\} = P\{\text{de seleção de uma moeda verdadeira}\} \cdot P\{\text{de obtenção de 2 caras com uma moeda verdadeira}\} + P\{\text{de seleção de uma moeda viciada}\} \cdot P\{\text{de obtenção de duas caras com moedas viciadas}\} = \frac{5}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{6} \cdot 1 \cdot 1 = \frac{3}{8}$.

23. Se A representa a produção de um item defeituoso e B simboliza a perda de um item defeituoso por parte da inspeção, então $P\{\text{de se comparar um item defeituoso}\} =$
$= P\{A\}\ P\{B \mid A\} = (0{,}04)\ (0{,}25) = 0{,}01.$

24. Se A simboliza a presença de chuva e B o ganho, então
$P\{B\} = P\{A\}\ P\{B \mid A\} + P\{\bar{A}\}\ P\{B \mid \bar{A}\} = (0{,}40)\ (0{,}5) +$
$+ (0{,}60)\ (0{,}7) = 0{,}62.$

25. (a) $\left(\frac{1}{2}\right)^5 = \frac{1}{32}$, (b) $\left(\frac{2}{3}\right)^5 = \frac{32}{243}$.

26. (a) $(0{,}04)\ (0{,}05) = 0{,}002$,
(b) $(0{,}04)\ (0{,}95) + (0{,}96)\ (0{,}05) = 0{,}086$.

27. A probabilidade de os dois motores do bimotor funcionar bem é de $(0{,}997)\ (0{,}997) = 0{,}994$; portanto a probabilidade de que pelo menos um não funcione será $= 1 - 0{,}994 = 0{,}006$. Portanto, o monomotor é mais seguro.

28. Esta probabilidade deve ser pelo menos tão grande quanto o p que satisfaz a equação $p^5 = 0{,}9$. A solução pode ser obtida por meio de logaritmo, por calculadoras eletrônicas ou pelo método das tentativas e erros. A solução é 0,979.

29. (a) 0,00763, (b) 0,00006,
(c) 0,98461, (d) 0,01539, (e) 0,02313.

30. (a) $(0{,}9)^6 = 0{,}53$, (b) $(0{,}8)^6 = 0{,}26$.

31. $P\{A_1 \text{ ou } A_2\} = \frac{16}{36} + \frac{6}{36} - \frac{2}{36} = \frac{5}{9}$.

32. Se A significa "ter mais de um acidente" e B "falha no teste", então

$$P\{A\,B\} = \frac{P\{A\}\,P\{B\,A\}}{P\{A\}\,P\{B\,A\} + P\{\bar{A}\}\,P\{B \mid \bar{A}\}}$$
$$= \frac{(0{,}10)\ (0{,}70)}{(0{,}10)\ (0{,}70) + (0{,}90)\ (0{,}20)} = 0{,}28.$$

33. (a) $\frac{1}{10}$, (b) $\frac{1}{2} \cdot \frac{1}{5} + \frac{1}{2} \cdot \frac{1}{25} = \frac{3}{25}$, (e) $\frac{1}{2} \cdot \frac{1}{5} \div \frac{3}{25} = \frac{5}{6}$.

34. Se A significa "ser um estudante satisfatório" e B, "passar no teste", então

$$P\{A \mid B\} = \frac{(0{,}70)\ (0{,}80)}{(0{,}70)\ (0{,}80) + (0{,}30)\ (0{,}40)} = \frac{14}{17}.$$

35. Se A significar "ser defeituoso" e B, "passar nas duas inspeções", então

$$P\{A \mid B\} = \frac{(0{,}10)\ (0{,}05)\ (0{,}04)}{(0{,}10)\ (0{,}05)\ (0{,}04) + 0{,}90(1)} = 0{,}00022.$$

36. (a) $\binom{13}{4} / \binom{52}{4} = \frac{11}{4165}$, (b) $\frac{44}{4165}$,

(c) $\binom{39}{4} / \binom{52}{4} = \frac{6.327}{20.825} = 0{,}30.$

37. $\binom{4}{1}\binom{48}{4} / \binom{52}{5} = \frac{3.243}{10.829} = 0{,}30.$

38. SS ou CS; por isso, $p \cdot p + (1 - p)(1 - p) = 1 - 2p + 2p^2$.

39. Se A significar "passar no teste" e B, "ser um bom trabalhador", então

$$P\{B\} = P\{A\}\,P\{B \mid A\} + P\{\bar{A}\}\,P\{B \mid \bar{A}\}.$$
$$= (0{,}60)\ (0{,}80) + (0{,}40)\ (0{,}30) = 0{,}60.$$

40. $\binom{40}{8} / \binom{50}{8} = 0{,}14.$

41. $$P\{5\,B \mid \geqslant 3\,B\} = \frac{P\{5\,B \text{ e } \geqslant 3\,B\}}{P\{\geqslant 3\,B\}} = \frac{P\{5\,B\}}{P\{\geqslant 3\,B\}}$$
$$= \frac{\binom{26}{5} / \binom{52}{5}}{\left[\binom{26}{5} + \binom{26}{4}\ \binom{26}{1} + \binom{26}{3}\ \binom{26}{2}\right] / \binom{52}{5}}$$
$$= \frac{\binom{26}{5}}{\binom{26}{5} + \binom{26}{4}\ \binom{26}{1} + \binom{26}{3}\ \binom{26}{2}} = 0{,}05.$$

42. (a) $\binom{4}{1}\binom{48}{12} / \binom{52}{13} = \frac{9.139}{20.825} = 0{,}44.$

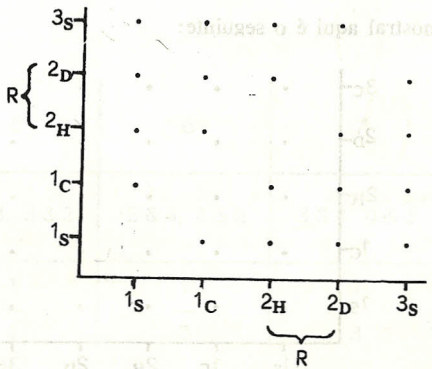
(b) $1 - \binom{48}{13} / \binom{52}{13} = 0{,}70.$

(c) $\binom{13}{13} / \binom{52}{13} = \frac{1}{6{,}35 \times 10^{11}}$; portanto mais ou menos uma chance em 600 000 000 000 mãos.

43. Uma vez que há seis vinhos numa linha e ele deve escolher três deles corretamente (digamos, Bordeaux), sua probabilidade de conseguir isto é

de $1/\binom{6}{3} = \frac{1}{20}$.

44. O espaço amostral aqui é o seguinte:



(a) $\frac{2}{10} = \frac{1}{10}$, (b) $\frac{4}{20} = \frac{1}{5}$, (c) $\frac{8}{20} = \frac{2}{5}$, (d) $\frac{4}{20} = \frac{1}{5}$,

(e) $\frac{6}{20} = \frac{3}{10}$, (f) $\frac{4}{6} = \frac{2}{3}$, (g) $\frac{1}{4}$.

Usando as regras para os eventos de dois estágios,

(a) $\frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$, (b) $\frac{2}{5} \cdot \frac{2}{4} = \frac{1}{5}$,

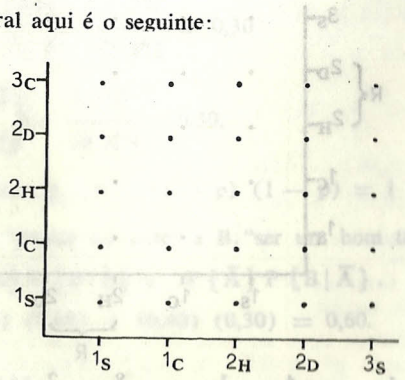
(c) $\frac{2}{5} \cdot \frac{1}{4} + \frac{3}{5} \cdot \frac{2}{4} = \frac{2}{5}$, (d) $\frac{2}{5} \cdot \frac{1}{4} + \frac{1}{5} \cdot \frac{2}{4}$,

(e) $\frac{6}{20} = \frac{3}{10}$, (f) $P\{A_1\} = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,

(f) $P\{A_1 \text{ e } A_2\} = \frac{4}{20} = \frac{1}{5}$, portanto, $P\{A_2 \mid A_1\} = \frac{\frac{1}{5}}{\frac{3}{10}} = \frac{2}{3}$,

(g) $P\{A_1 \mid A_2\} = \frac{\frac{1}{5} \cdot \frac{1}{4}}{\frac{1}{5}\left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right)} = \frac{1}{4}$.

45. O espaço amostral aqui é o seguinte:



(a) $\frac{4}{25}$, (b) $\frac{9}{25}$, (c) $\frac{16}{25}$, (d) 1, (e) $\frac{12}{25}$, (f) $\frac{12}{16}$.

Usando as regras para os eventos de dois estágios,

(a) $\frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25}$, (b) $\frac{1}{5} \cdot \frac{1}{5} + \frac{1}{5} \cdot \frac{4}{5} + \frac{4}{5} \cdot \frac{1}{5} = \frac{9}{25}$,

(c) $1 - \frac{3}{5} \cdot \frac{3}{5} = \frac{16}{25}$, (d) 1,

(e) $\frac{2}{5} \cdot \frac{2}{5} + \frac{2}{5} \cdot \frac{2}{5} + \frac{2}{5} \cdot \frac{2}{5} = \frac{12}{25}$, (f) $\frac{12/25}{16/25} = \frac{12}{16}$.

CAPÍTULO 4

3.

x	4	5	6	7	8
Eventos	2 & 2	2 & 3, 3 & 2	3 & 3, 2 & 4, 4 & 2	3 & 4, 4 & 3	4 & 4
P{x}	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

4.

x	5	6	7
Eventos	2 & 3, 3 & 2	2 & 4, 4 & 2	3 & 4, 4 & 3
P{x}	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

5.

x	0	1	2	3
Sucessos	Nenhum	1, 2, 3	1 & 2, 1 & 3, 2 & 3	1 & 2 & 3
P{x}	$\left(\frac{3}{4}\right)^3$	$3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2$	$3\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)$	$\left(\frac{1}{4}\right)^3$

6. $\mu = 4 \cdot \frac{1}{9} + 5 \cdot \frac{2}{9} + 6 \cdot \frac{3}{9} + 7 \cdot \frac{2}{9} + 8 \cdot \frac{1}{9} = 6$ ou por simetria.

$\sigma^2 = 4 \cdot \frac{1}{9} + 1 \cdot \frac{2}{9} + 1 \cdot \frac{2}{9} + 4 \cdot \frac{1}{9} = \frac{12}{9}$, $\sigma = \sqrt{4/3} = 1{,}15$.

7. $\mu = 5 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} + 7 \cdot \frac{1}{3} = 6$, ou por simetria.

$\sigma^2 = 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}$, $\sigma = \sqrt{2/3} = 0{,}82$.

8. $\mu = 0 \cdot \frac{1}{10} + 1 \cdot \frac{1}{10} + \ldots + 9 \cdot \frac{1}{10} = 4{,}5$, ou por simetria.

$\sigma^2 = 2\left[\left(\frac{9}{2}\right)^2 + \left(\frac{7}{2}\right)^2 + \left(\frac{5}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]\frac{1}{10} = \frac{33}{4}$,

$\sigma = \sqrt{33/4} = 2{,}87$.

9. $\mu = 0$, $\sigma = \sqrt{3/4} = 0{,}87$; portanto, 25%.

10. $\mu = 1 \cdot \frac{27}{64} + 2 \cdot \frac{9}{64} + 3 \cdot \frac{1}{64} = \frac{3}{4}$.

$\sigma^2 = \frac{9}{16} \cdot \frac{27}{64} + \frac{1}{16} \cdot \frac{27}{64} + \frac{25}{16} \cdot \frac{9}{64} \cdot \frac{81}{16} \cdot \frac{1}{64} = \frac{9}{16}$, $\sigma = \frac{3}{4}$.

11. \$ 205.

12. A probabilidade de se obter pelo menos duas caras é de 1/2 e a de se obter qualquer número particular de pontos é de 1/6; portanto

$$E(x) = 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + \ldots + 6 \cdot \frac{1}{12} = \frac{21}{12} = 1{,}75.$$

13. É necessário resolver a seguinte equação para p: $6.000p + 7.000(1-p) = 2.000p + 10.000(1-p)$.

A solução é $p = \frac{3}{7}$.

14. $E(I) = 700(0{,}9) + 600(0{,}1) = 690$, $E(N) = 750(0{,}5) + 650(0{,}5) = 700$; portanto, não fazer publicidade.

15

x	2	14	5	10	15
P{x}	$\frac{1}{8}\cdot\frac{1}{2}$	$\frac{1}{8}\cdot\frac{1}{2}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

; portanto $E[x] = \$\ 8{,}50$.

16.

(a)

x	0	1
P{x}	$\frac{4}{9}$	$\frac{5}{9}$

(b)

x	3	4	5
P{x}	$\frac{3}{9}$	$\frac{4}{9}$	$\frac{2}{9}$

(c)

x	0	1	2
P{x}	$\frac{4}{9}\cdot\frac{4}{9}$	$\frac{5}{9}\cdot\frac{4}{9}+\frac{4}{9}\cdot\frac{5}{9}$	$\frac{5}{9}\cdot\frac{5}{9}$

(d)

x	0	1	2
P{x}	$\frac{4}{9}\cdot\frac{3}{8}$	$\frac{5}{9}\cdot\frac{4}{8}+\frac{4}{9}\cdot\frac{5}{8}$	$\frac{5}{9}\cdot\frac{4}{8}$

17. $E(D) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} +$

$+ 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} +$

$+ 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$, ou por simetria.

$E(C) = 2\left[1 \cdot \frac{4}{16} + 2 \cdot \frac{6}{16} + 3 \cdot \frac{4}{16} + 4 \cdot \frac{1}{16}\right] = 4$;

portanto, escolha o dado.

18.

x	1	2	3	4	5	6
P{x}	$\frac{3}{8}+\frac{1}{8}\cdot\frac{1}{6}$	$\frac{3}{8}+\frac{1}{8}\cdot\frac{1}{6}$	$\frac{1}{8}\cdot\frac{1}{6}$	$\frac{1}{8}\cdot\frac{1}{6}$	$\frac{1}{8}\cdot\frac{1}{6}$	$\frac{1}{8}\cdot\frac{1}{6}$

$E(x) = 1 \cdot \frac{19}{48} + 2 \cdot \frac{19}{48} + 3 \cdot \frac{1}{48} + 4 \cdot \frac{1}{48} +$

$$+ 5 \cdot \frac{1}{48} + 6 \cdot \frac{1}{48} = \frac{75}{48} = 1{,}56.$$

CAPÍTULO 5

2.

x	0	1	2	3	4
P{x}	$\left(\frac{1}{2}\right)^4$	$4\left(\frac{1}{2}\right)^4$	$6\left(\frac{1}{2}\right)^4$	$4\left(\frac{1}{2}\right)^4$	$\left(\frac{1}{2}\right)^4$

3.

x	0	1	2	3	4
P{x}	$\left(\frac{5}{6}\right)^4$	$4\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)$	$6\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)^2$	$4\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^3$	$\left(\frac{1}{6}\right)^4$

4.

x	0	1	2	3	4	5
P{x}	$\left(\frac{1}{2}\right)^5$	$5\left(\frac{1}{2}\right)^5$	$10\left(\frac{1}{2}\right)^5$	$10\left(\frac{1}{2}\right)^5$	$5\left(\frac{1}{2}\right)^5$	$\left(\frac{1}{2}\right)^5$

5.

x	0	1	2	3
P{x}	$\left(\frac{3}{4}\right)^5$	$5\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^4$	$10\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3$	$10\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^2$

4	5
$5\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)$	$\left(\frac{1}{4}\right)^5$

6. $P\{x \geqslant 6\} = \sum_{x=6}^{8} \frac{8!}{x!(8-x)!}(0{,}8)^x\,(0{,}2)^{8-x} = 0{,}797.$

7. $P\{x \geqslant 8\} = \sum_{x=8}^{10} \frac{10!}{x!(10-x)!}\left(\frac{1}{2}\right)^{10} = 0{,}055.$

8. $P\{x \geqslant 9\} = \sum_{x=9}^{12} \frac{12!}{x!(12-x)!}\left(\frac{1}{2}\right)^{12} = \frac{299}{4\,096} = 0{,}073.$

9. $\sum_{x=6}^{n} \frac{n!}{x!(n-x)!}(0{,}8)^x\,(0{,}2)^{n-x} \geqslant 0{,}9.$ Tente diferentes valores de n começando por 9. Da Tabela III e $n = 9$, $P\{x \geqslant 6\} = 1 - 0{,}0856 = 0{,}9154$; portanto $n = 9$ deveria bastar.

10. Os aumentos e as quedas, de dia para dia, não são eventos independentes.

11. A chuva freqüentemente dura mais que um dia, de modo que os dias consecutivos não constituem eventos independentes com relação à chuva.

12. $\mu = 8 \cdot \frac{1}{4} = 2,\ \sigma = \sqrt{8 \cdot \frac{1}{4} \cdot \frac{3}{4}} = \sqrt{\frac{3}{2}} = 1{,}22.$

13. $\mu = \frac{5}{4},\ \sigma = \sqrt{5 \cdot \frac{1}{4} \cdot \frac{3}{4}} = \frac{\sqrt{15}}{4} = 0{,}97.$

16. (a) $z = \frac{14 - 12}{2} = 1,\ P\{z > 1\} = 0{,}16.$

(b) $z = \frac{11 - 12}{2} = -\frac{1}{2},\ P\{z > -0{,}5\} = 0{,}5 + P\{0 < z < 0{,}5\} = 0{,}69.$

(c) $z = \frac{10 - 12}{2} = -1,\ P\{z < -1\} = P\{z > 1\} = 0{,}16.$

(d) $z = \frac{10{,}5 - 12}{2} = -0{,}75,\ P\{z < -0{,}75\} = P\{z > 0{,}75\} = 0{,}23.$

(e) $z_1 = \frac{10 - 12}{2} = -1,\ z_2 = \frac{13 - 12}{2} = \frac{1}{2},\ P\{-1 < z < 0{,}5\} = P\{0 < z < 0{,}5\} + P\{0 < z < 1\} = 0{,}19 + 0{,}34 = 0{,}53.$

17. (a) $z = \frac{66 - 69}{3} = -1,\ P\{z < -1\} = P\{z > 1\} = 0{,}16.$

(b) $z_1 = \frac{65 - 69}{3} = -1{,}33,\ z_2 = \frac{71 - 69}{3} = 0{,}67,$

$P\{-1{,}33 < z < 0{,}67\} = P\{0 < z < 0{,}67\} + P\{0 < z < 1{,}33\} = 0{,}25 + 0{,}41 = 0{,}66.$

18. $P\{z > 0,8\} = 0,21.$

19. $z = \dfrac{x - 58}{4}$, $P\{z > 0,84\} = 0,20$; portanto, x deve ser pelo menos tão grande quanto a solução para $\dfrac{x - 58}{4} = 0,84$, que é 61,4 polegadas ou 5 pés e 1,4 de polegada.

20. (a) $z = \dfrac{130 - 115}{8} = 1,88$, $P\{z > 1,88\} = 0,03.$

(b) $z = \dfrac{100 - 115}{8} = -1,88$, $P\{z < -1,88\} = 0,03.$

(c) $z_1 = \dfrac{105 - 115}{8} = -1,25$, $z_2 = \dfrac{125 - 115}{8} = 1,25,$

$P\{-1,25 < z < 1,25\} = 2P\{0 < z < 1,25\} = 0,79.$

21. $z = \dfrac{100 - 130}{20} = -1,5$, $P\{z < -1,5\} = 0,07.$

22. $z = \dfrac{x-6}{2}$, $P\{z < -1,04\} = 0,15$; portanto x deve ser no máximo tão grande quanto a solução para $\dfrac{x-6}{2} = -1,04$, que é 3,92. Uma garantia de cerca de 4 anos deveria bastar.

23. (a) $P\{x = 6\} = \dfrac{8!}{6!\,2!}\left(\dfrac{1}{2}\right)^8 = \dfrac{28}{256} = 0,109,$

$z_1 = \dfrac{5,5 - 4}{1,41} = 1,06$, $z_2 = \dfrac{6,5 - 4}{1,41} = 1,77,$

$P\{1,06 < z < 1,77\} = 0,4616 - 0,3554 = 0,106.$

(b) Da Tabela III, $P\{x \geqslant 6\} = 0,1445,$

$P\{z > 1,06\} = 0,1446.$

24. Da Tabela III, $P\{x \geqslant 4\} = 0,710,$

$z = \dfrac{3,5 - 4,2}{1,30} = -0,54$, $P\{z > -0,54\} = 0,705.$

25. $z = \dfrac{8,5 - 5,4}{1,94} = 1,60$, $P\{z > 1,60\} = 0,05.$

26. (a) $z = \dfrac{13,5 - 10}{3} = 1,17$, $P\{z > 1,17\} = 0,12.$

(b) $z_1 = \dfrac{4,5 - 10}{3} = -1,83$, $z_2 = \dfrac{14,5 - 10}{3} = 1,50,$

$P\{-1,83 < z < 1,50\} = 0,90.$

27. $z = \dfrac{0,25 - 0,20}{\sqrt{\dfrac{(0,2)\,(0,8)}{200}}} = \dfrac{0,05}{0,028} = 1,79$, $P\{z > 1,79\} = 0,04.$

28. $z = \dfrac{0,15 - 0,10}{\sqrt{\dfrac{(0,1)\,(0,9)}{200}}} = \dfrac{0,05}{0,021} = 2,38$, $P\{z > 2,38\} = 0,01.$

29. $z = \dfrac{7,5 - 5}{1,94} = 1,29$, $P\{z > 1,29\} = 0,10.$

30. $n = 20$, $p = 0,2$, $z = \dfrac{5,5 - 4}{1,79} = 0,84$, $P\{z < 0,84\} = 0,80.$

31. $P\{0\} = \left(\dfrac{9}{10}\right)^{20} = 0,12.$ Como 0 é o menor valor possível de x e sua probabilidade é tão grande, não devemos esperar que a cauda esquerda da curva normal se ajuste muito bem a esta distribuição binomial.

32. $\mu = 2$, $\sigma = \sqrt{20 \cdot \dfrac{1}{10} \cdot \dfrac{9}{10}} = 1,34$, $z_1 = \dfrac{-0,5 - 2}{1,34} = -1,87,$

$z_2 = \dfrac{0,5 - 2}{1,34} = -1,12$, $P\{-1,87 < z < -1,12\} =$

$= 0,4693 - 0,3686 = 0,10.$ Surpreendentemente próxima.

34. $P\{x \geqslant 5\} = \sum_{x=5}^{8} \dfrac{8!}{x!(8-x)!}\left(\dfrac{1}{2}\right)^8 = 0,36.$

35. (a) $P\{x\} = \dfrac{10!}{x!(10 - x)!}\left(\dfrac{5}{9}\right)^x \left(\dfrac{4}{9}\right)^{10-x}$

(b) $P\{4 \leqslant x \leqslant 5\} = \sum_{x=4}^{5} \frac{10!}{x!(10-x)!} \left(\frac{5}{9}\right)^x \left(\frac{4}{9}\right)^{10-x} =$

$= 210 \left(\frac{5}{9}\right)^4 \left(\frac{4}{9}\right)^6 + 252 \left(\frac{5}{9}\right)^5 \left(\frac{4}{9}\right)^5 = 0{,}39.$

(c) $z_1 = \frac{3{,}5 - 5{,}56}{1{,}57} = -1{,}31, \quad z_2 = \frac{5{,}5 - 5{,}56}{1{,}57} = -0{,}04,$

$P\{-1{,}31 < z < -0{,}04\} = 0{,}39.$

(d) $z = \frac{0{,}60 - 0{,}556}{\sqrt{\frac{\frac{5}{9} \cdot \frac{4}{9}}{100}}} = 0{,}88, \quad P\{z < 0{,}88\} = 0{,}81.$

36. (a) $z = \frac{11{,}9 - 12{,}1}{0{,}1} = -2, \quad P\{z < -2\} = 0{,}02.$

(b) $z = \frac{11{,}9 - \mu}{0{,}1}$, $P\{z < -1{,}64\} = 0{,}05$; portanto μ deve ser pelo menos tão grande quanto a solução da equação $\frac{11{,}9 - \mu}{0{,}1} = -1{,}64$.

A solução é $\mu = 12{,}064$ ou $12{,}1$. O resultado atual é, portanto, mais que satisfatório.

37. Uma vez que todos os passageiros de primeira classe podem estar bem acomodados, o problema se resume em calcular $P\{x \leqslant 300\}$. Aqui

$z = \frac{300{,}5 - 288}{5{,}37} = 2{,}33, \quad P\{z \leqslant 2{,}33\} = 0{,}99.$

CAPÍTULO 6

1. Através das matrículas e números aleatórios, ou, digamos, escolha vinte classes das 10 horas, aleatoriamente, e selecione, ao acaso, cinco estudantes de cada classe.
2. Provavelmente a letra "A" estará mais ligada a nomes de origem inglesa e escandinava; portanto, se tais indivíduos tenderem a ter rendas mais altas ou mais baixas que a média, uma amostra dos "A" não seria satisfatória.
3. As pessoas que respondem, geralmente ou são muito bondosas ou são pessoas difíceis de satisfazer. A maioria das pessoas não se dará ao trabalho de responder. Portanto, a amostra refletirá apenas as posições extremadas.
4. A maior parte das pessoas não escreve aos congressistas, mesmo se tiverem opiniões bastante definidas sobre determinado assunto. Os grupos de pressão, freqüentemente, pedem que seus membros escrevam. Os pontos de vista do cidadão comum raramente são ouvidos.
5. Seria preciso descobrir se as 600 que não responderam têm os mesmos pontos de vista que as que responderam, antes de podermos confiar nos dados obtidos das 400.
6. Escolha dois números aleatórios de três dígitos e use-os como coordenadas de um ponto nos 1 000 pés quadrados sobre um lado que tenha dois destes lados sobre os eixos positivos de x e y. Isto localiza o ponto aleatório sobre o qual centra-se o arco.
7. As unidades familiais diferem em tamanho com relação aos adultos; portanto, as unidades familiais com muitos adultos não estarão representadas convenientemente.
8. (a) Suponha uma floresta retangular. Selecione aleatoriamente 10 pontos ao longo da base do retângulo e, em cada ponto, trace uma perpendicular à linha da base. Escolha 10 pontos aleatórios ao longo de cada linha vertical dentro do retângulo. Escolha a árvore mais próxima (centro da árvore) para cada ponto.

9. (a) $z = \frac{21 - 20}{0{,}5} = 2, \quad P\{z > 2\} = 0{,}02.$

(b) $z = \frac{19{,}5 - 20}{0{,}5} = -1, \quad P\{z > -1\} = 0{,}84.$

(c) $z_1 = \frac{19 - 20}{0{,}5} = -2, \quad z_2 = \frac{21 - 20}{0{,}5} = 2,$

$P\{-2 < z < 2\} = 0{,}95.$

(d) $z = \frac{22 - 20}{0{,}5} = 4, \quad P\{z \geqslant 4\} = 0{,}00.$

10. (a) $z = \frac{26 - 25}{2} = 0{,}50, \quad P\{z < 0{,}50\} = 0{,}69.$

(b) $z = \frac{31 - 25}{2} = 3{,}00, \quad P\{z > 3{,}00\} = 0{,}00.$

(c) $z = \frac{24 - 25}{2} = -0{,}50, \quad P\{z > -0{,}50\} = 0{,}69.$

(d) $z = \dfrac{21 - 25}{2} = -2{,}00$, $P\{z < -2{,}00\} = 0{,}02$.

(e) $z_1 = \dfrac{28 - 25}{2} = 1{,}50$, $z_2 = \dfrac{29 - 25}{2} = 2{,}00$,

$P\{1{,}50 < z < 2{,}00\} = 0{,}4772 - 0{,}4332 = 0{,}04$.

11. A segunda curva deveria ser três vezes mais alta que a primeira em $x = 10$ e ter um terço da largura.

12. Seria aproximadamente o dobro da altura e a metade da largura da curva $\bar{x}$.

13. $z = \dfrac{\bar{x} - \mu}{0{,}7} = \dfrac{1}{0{,}7} = 1{,}43$, $P\{|z| > 1{,}43\} = 0{,}15$.

14. $z = \dfrac{\bar{x} - \mu}{0{,}4} = \dfrac{0{,}5}{0{,}4} = 1{,}25$, $P\{|z| > 1{,}25\} = 0{,}21$.

15. $z = \dfrac{159 - 154}{2} = 2{,}50$, $P\{z > 2{,}50\} = 0{,}01$;

Portanto, parecem ser mais pesados, ou os primeiros 100 estudantes matriculados podem não ser típicos.

19. Como $x = 0$ ou 1, a proporção amostral $\dfrac{x}{n}$ é uma média amostral dos valores observados de x. Porque $\mu = E(x) = 0.q+1.p = p$ e $\sigma^2 = (0-p)^2\, q+(1-p)^2\, p = pq$, segue-se do Teorema 2 que $\dfrac{x}{n}$ terá uma distribuição normal aproximada com média p e desvio-padrão $\sigma/\sqrt{n} = \sqrt{pq/n}$, desde que n seja suficientemente grande. O que justifica a aproximação normal na fórmula (7), Capítulo 5.º.

CAPÍTULO 7

1. (a) $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{16}} = 3$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(3) = 5{,}88$.

(b) $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{64}} = 1{,}50$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(1{,}50) = 2{,}94$. Obter-se-ia uma precisão duas vezes maior.

2. $\sigma_{\bar{x}} = \dfrac{2}{\sqrt{25}} = 0{,}4$; portanto existe uma probabilidade de 0,95 de que o erro não exceda $1{,}96(0{,}4) = 0{,}78$.

(b) Precisão duas vezes maior.

3. Usando $z_0 = 1{,}96$, (a) $n = \left(\dfrac{1{,}96 \cdot 12}{3}\right)^2 \cong 62$.

(b) $n = \left(\dfrac{1{,}96 \cdot 12}{1}\right)^2 \cong 553$.

4. Usando $z_0 = 1{,}64$, (a) $n = \left(\dfrac{1{,}64 \cdot 12}{3}\right)^2 \cong 43$.

(b) $n = \left(\dfrac{1{,}64 \cdot 12}{1}\right)^2 \cong 387$.

5. (a) $n = \left(\dfrac{1{,}96 \cdot 2}{1/2}\right)^2 \cong 62$, (b) $n = \left(\dfrac{1{,}96 \cdot 2}{1/5}\right)^2 \cong 384$.

6. $n = \left(\dfrac{1{,}64 \cdot 2}{1/2}\right)^2 = 43$.

7. (a) $140 - 1{,}96\dfrac{12}{\sqrt{16}} < \mu < 140 + 1{,}96\dfrac{12}{\sqrt{16}}$; portanto,

$134{,}1 < \mu < 145{,}9$.

(b) $140 - 1{,}64\dfrac{12}{\sqrt{16}} < \mu < 140 + 1{,}64\dfrac{12}{\sqrt{16}}$; portanto,

$135{,}1 < \mu < 144{,}9$.

8. (a) $54 - 1{,}96\dfrac{2}{\sqrt{25}} < \mu < 54 + 1{,}96\dfrac{2}{\sqrt{25}}$; portanto,

$53{,}2 < \mu < 54{,}8$.

(b) $54 - 1{,}64\dfrac{2}{\sqrt{25}} < \mu < 54 + 1{,}64\dfrac{2}{\sqrt{25}}$; portanto,

$53{,}3 < \mu < 54{,}7$.

9. (a) $\sigma_{\bar{x}} \cong \frac{4}{\sqrt{60}} = 0{,}52$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(0{,}52) = 1{,}02$.

(b) $n = \left(\frac{1{,}96 \cdot 4}{1/2}\right)^2 \cong 246$.

(c) $42 - 1{,}96 \frac{4}{\sqrt{60}} < \mu < 42 + 1{,}96 \frac{4}{\sqrt{60}}$; portanto,

$41{,}0 < \mu < 43{,}0$.

10. $\sigma_{\bar{x}} \cong 0{,}6$, $z = \frac{1}{0{,}6} = 1{,}67$; portanto, $P\{|z| < 1{,}67\} = 0{,}90$.

11. $19{,}6 - 1{,}64 \frac{0{,}7}{\sqrt{30}} < \mu < 19{,}6 + 1{,}64 \frac{0{,}7}{\sqrt{30}}$; portanto,

$19{,}4 < \mu < 19{,}8$.

12. $n = \left(\frac{1{,}96 \cdot 2}{0{,}64}\right)^2 \cong 38$.

13. Como $n \cong 38$, baseado em $\sigma = 2$, tome uma amostra, por exemplo de 25; calcule o valor do s; use s em vez de σ para obter uma nova estimativa de n; tome tantas amostras adicionais quantas necessário para totalizar o segundo valor de n. Para maior precisão compare σ e s e use só a metade do tamanho amostral mínimo adicional necessário para repetir as estimativas de n.

14. $\sigma_{\hat{p}} = \sqrt{\frac{(0{,}6)\ (0{,}4)}{300}} = 0{,}028$; portanto, há uma probabilidade de 0,95 de que o erro de estimativa seja menor do que $1{,}96(0{,}028) = 0{,}055$.

15. $n = \frac{(0{,}6)\ (0{,}4)\ (1{,}96)^2}{(0{,}04)^2} \cong 576$.

16. $n = \frac{(0{,}03)\ (0{,}97)\ (2{,}17)^2}{(0{,}005)^2} \cong 5.481$.

17. $n = \frac{(0{,}5)\ (0{,}5)\ (1{,}96)^2}{(0{,}04)^2} \cong 600$.

18. $n = \frac{(0{,}5)\ (0{,}5)\ (1{,}64)^2}{(0{,}03)^2} \cong 747$.

19. $0{,}70 - 1{,}96 \sqrt{\frac{(0{,}7)\ (0{,}3)}{400}} < p < 0{,}70 + 1{,}96 \sqrt{\frac{(0{,}7)\ (0{,}3)}{400}}$;

portanto, $0{,}66 < p < 0{,}74$.

20. $0{,}40 - 1{,}96 \sqrt{\frac{(0{,}4)\ (0{,}6)}{2\ 000}} < p < 0{,}40 + 1{,}96 \sqrt{\frac{(0{,}4)\ (0{,}6)}{2\ 000}}$;

portanto, $0{,}38 < p < 0{,}42$.

21. (a) Os primeiros estudantes a chegar podem não ser típicos.

(b) $0{,}30 - 2{,}05 \sqrt{\frac{(0{,}3)\ (0{,}7)}{200}} < p < 0{,}30 + 2{,}05 \sqrt{\frac{(0{,}3)\ (0{,}7)}{200}}$;

portanto, $0{,}23 < p < 0{,}37$.

22. $0{,}07 - 1{,}64 \sqrt{\frac{(0{,}07)\ (0{,}93)}{7\ 000}} < p < 0{,}07 + 1{,}64 \sqrt{\frac{(0{,}07)\ (0{,}93)}{7\ 000}}$,

ou $0{,}065 < p < 0{,}075$; portanto, $7\ 000\ (0{,}065) < 7\ 000\ p < 7\ 000\ (0{,}075)$, ou 455< número de acidentes < 525.

24. $20 - 2{,}201 \frac{4}{\sqrt{12}} < \mu < 20 + 2{,}201 \frac{4}{\sqrt{12}}$;

portanto, $17{,}5 < \mu < 22{,}5$.

25. $22 - 2{,}093 \frac{4}{\sqrt{20}} < \mu < 22 + 2{,}093 \frac{4}{\sqrt{20}}$;

portanto, $20{,}1 < \mu < 23{,}9$.

26. $22 - 1{,}96 \frac{4}{\sqrt{20}} < \mu < 22 + 1{,}96 \frac{4}{\sqrt{20}}$;

portanto, $20{,}2 < \mu < 23{,}8$.

(d) $z = \dfrac{21 - 25}{2} = -2{,}00$, $P\{z < -2{,}00\} = 0{,}02$.

(e) $z_1 = \dfrac{28 - 25}{2} = 1{,}50$, $z_2 = \dfrac{29 - 25}{2} = 2{,}00$,

$P\{1{,}50 < z < 2{,}00\} = 0{,}4772 - 0{,}4332 = 0{,}04$.

11. A segunda curva deveria ser três vezes mais alta que a primeira em $x = 10$ e ter um terço da largura.

12. Seria aproximadamente o dobro da altura e a metade da largura da curva $\bar{x}$.

13. $z = \dfrac{\bar{x} - \mu}{0{,}7} = \dfrac{1}{0{,}7} = 1{,}43$, $P\{|z| > 1{,}43\} = 0{,}15$.

14. $z = \dfrac{\bar{x} - \mu}{0{,}4} = \dfrac{0{,}5}{0{,}4} = 1{,}25$, $P\{|z| > 1{,}25\} = 0{,}21$.

15. $z = \dfrac{159 - 154}{2} = 2{,}50$, $P\{z > 2{,}50\} = 0{,}01$;

Portanto, parecem ser mais pesados, ou os primeiros 100 estudantes matriculados podem não ser típicos.

19. Como $x = 0$ ou 1, a proporção amostral $\dfrac{x}{n}$ é uma média amostral dos valores observados de x. Porque $\mu = E(x) = 0.q+1.p = p$ e $\sigma^2 = (0-p)^2\, q+(1-p)^2\, p = pq$, segue-se do Teorema 2 que $\dfrac{x}{n}$ terá uma distribuição normal aproximada com média p e desvio-padrão $\sigma/\sqrt{n} = \sqrt{pq/n}$, desde que n seja suficientemente grande. O que justifica a aproximação normal na fórmula (7), Capítulo 5.º.

CAPÍTULO 7

1. (a) $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{16}} = 3$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(3) = 5{,}88$.

(b) $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{64}} = 1{,}50$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(1{,}50) = 2{,}94$. Obter-se-ia uma precisão duas vezes maior.

2. $\sigma_{\bar{x}} = \dfrac{2}{\sqrt{25}} = 0{,}4$; portanto existe uma probabilidade de 0,95 de que o erro não exceda $1{,}96(0{,}4) = 0{,}78$.

(b) Precisão duas vezes maior.

3. Usando $z_0 = 1{,}96$, (a) $n = \left(\dfrac{1{,}96 \cdot 12}{3}\right)^2 \cong 62$.

(b) $n = \left(\dfrac{1{,}96 \cdot 12}{1}\right)^2 \cong 553$.

4. Usando $z_0 = 1{,}64$, (a) $n = \left(\dfrac{1{,}64 \cdot 12}{3}\right)^2 \cong 43$.

(b) $n = \left(\dfrac{1{,}64 \cdot 12}{1}\right)^2 \cong 387$.

5. (a) $n = \left(\dfrac{1{,}96 \cdot 2}{1/2}\right)^2 \cong 62$, (b) $n = \left(\dfrac{1{,}96 \cdot 2}{1/5}\right)^2 \cong 384$.

6. $n = \left(\dfrac{1{,}64 \cdot 2}{1/2}\right)^2 = 43$.

7. (a) $140 - 1{,}96\dfrac{12}{\sqrt{16}} < \mu < 140 + 1{,}96\dfrac{12}{\sqrt{16}}$; portanto,

$134{,}1 < \mu < 145{,}9$.

(b) $140 - 1{,}64\dfrac{12}{\sqrt{16}} < \mu < 140 + 1{,}64\dfrac{12}{\sqrt{16}}$; portanto,

$135{,}1 < \mu < 144{,}9$.

8. (a) $54 - 1{,}96\dfrac{2}{\sqrt{25}} < \mu < 54 + 1{,}96\dfrac{2}{\sqrt{25}}$; portanto,

$53{,}2 < \mu < 54{,}8$.

(b) $54 - 1{,}64\dfrac{2}{\sqrt{25}} < \mu < 54 + 1{,}64\dfrac{2}{\sqrt{25}}$; portanto,

$53{,}3 < \mu < 54{,}7$.

9. (a) $\sigma_{\bar{x}} \cong \frac{4}{\sqrt{60}} = 0{,}52$; portanto, há uma probabilidade de 0,95 de que o erro não exceda $1{,}96(0{,}52) = 1{,}02$.

(b) $n = \left(\frac{1{,}96 \cdot 4}{1/2}\right)^2 \cong 246$.

(c) $42 - 1{,}96 \frac{4}{\sqrt{60}} < \mu < 42 + 1{,}96 \frac{4}{\sqrt{60}}$; portanto,

$41{,}0 < \mu < 43{,}0$.

10. $\sigma_{\bar{x}} \cong 0{,}6$, $z = \frac{1}{0{,}6} = 1{,}67$; portanto, $P\{|z| < 1{,}67\} = 0{,}90$.

11. $19{,}6 - 1{,}64 \frac{0{,}7}{\sqrt{30}} < \mu < 19{,}6 + 1{,}64 \frac{0{,}7}{\sqrt{30}}$; portanto,

$19{,}4 < \mu < 19{,}8$.

12. $n = \left(\frac{1{,}96 \cdot 2}{0{,}64}\right)^2 \cong 38$.

13. Como $n \cong 38$, baseado em $\sigma = 2$, tome uma amostra, por exemplo de 25; calcule o valor do s; use s em vez de σ para obter uma nova estimativa de n; tome tantas amostras adicionais quantas necessário para totalizar o segundo valor de n. Para maior precisão compare σ e s e use só a metade do tamanho amostral mínimo adicional necessário para repetir as estimativas de n.

14. $\sigma_{\hat{p}} = \sqrt{\frac{(0{,}6)\ (0{,}4)}{300}} = 0{,}028$; portanto, há uma probabilidade de 0,95 de que o erro de estimativa seja menor do que $1{,}96(0{,}028) = 0{,}055$.

15. $n = \frac{(0{,}6)\ (0{,}4)\ (1{,}96)^2}{(0{,}04)^2} \cong 576$.

16. $n = \frac{(0{,}03)\ (0{,}97)\ (2{,}17)^2}{(0{,}005)^2} \cong 5.481$.

17. $n = \frac{(0{,}5)\ (0{,}5)\ (1{,}96)^2}{(0{,}04)^2} \cong 600$.

18. $n = \frac{(0{,}5)\ (0{,}5)\ (1{,}64)^2}{(0{,}03)^2} \cong 747$.

19. $0{,}70 - 1{,}96 \sqrt{\frac{(0{,}7)\ (0{,}3)}{400}} < p < 0{,}70 + 1{,}96 \sqrt{\frac{(0{,}7)\ (0{,}3)}{400}}$;

portanto, $0{,}66 < p < 0{,}74$.

20. $0{,}40 - 1{,}96 \sqrt{\frac{(0{,}4)\ (0{,}6)}{2\ 000}} < p < 0{,}40 + 1{,}96 \sqrt{\frac{(0{,}4)\ (0{,}6)}{2\ 000}}$;

portanto, $0{,}38 < p < 0{,}42$.

21. (a) Os primeiros estudantes a chegar podem não ser típicos.

(b) $0{,}30 - 2{,}05 \sqrt{\frac{(0{,}3)\ (0{,}7)}{200}} < p < 0{,}30 + 2{,}05 \sqrt{\frac{(0{,}3)\ (0{,}7)}{200}}$;

portanto, $0{,}23 < p < 0{,}37$.

22. $0{,}07 - 1{,}64 \sqrt{\frac{(0{,}07)\ (0{,}93)}{7\ 000}} < p < 0{,}07 + 1{,}64 \sqrt{\frac{(0{,}07)\ (0{,}93)}{7\ 000}}$,

ou $0{,}065 < p < 0{,}075$; portanto, $7\ 000\ (0{,}065) < 7\ 000\ p < 7\ 000\ (0{,}075)$, ou 455< número de acidentes < 525.

24. $20 - 2{,}201 \frac{4}{\sqrt{12}} < \mu < 20 + 2{,}201 \frac{4}{\sqrt{12}}$;

portanto, $17{,}5 < \mu < 22{,}5$.

25. $22 - 2{,}093 \frac{4}{\sqrt{20}} < \mu < 22 + 2{,}093 \frac{4}{\sqrt{20}}$;

portanto, $20{,}1 < \mu < 23{,}9$.

26. $22 - 1{,}96 \frac{4}{\sqrt{20}} < \mu < 22 + 1{,}96 \frac{4}{\sqrt{20}}$;

portanto, $20{,}2 < \mu < 23{,}8$.

27. $30{,}6 - 1{,}796 \dfrac{5{,}09}{\sqrt{12}} < \mu < 30{,}6 + 1{,}796 \dfrac{5{,}09}{\sqrt{12}}$;

portanto, $28{,}0 < \mu < 33{,}2$.

28. $32 - 1{,}729 \dfrac{18}{\sqrt{20}} < \mu < 32 + 1{,}729 \dfrac{18}{\sqrt{20}}$;

portanto, $25{,}0 < \mu < 39{,}0$.

29. $32 - 1{,}833 \dfrac{18}{\sqrt{10}} < \mu < 32 + 1{,}833 \dfrac{18}{\sqrt{10}}$, ou

$32 - 10{,}4 < \mu < 32 + 10{,}4$.

$32 - 1{,}684 \dfrac{18}{\sqrt{40}} < \mu < 32 + 1{,}684 \dfrac{18}{\sqrt{40}}$, ou

$32 - 4{,}8 < \mu < 32 + 4{,}8$.

O intervalo para $n = 10$ é duas vezes mais longo que para $n = 40$. Para o método da grande amostra deveria ser exatamente duas vezes mais longo.

30. (a)

x	0,5	1,5	2,5	3,5	4,5	6	8
f	17	42	205	105	19	9	4

$\bar{x} = 2{,}80$, $s = 1{,}10$.

(b) $2{,}80 - 1{,}96 \dfrac{1{,}10}{\sqrt{401}} < \mu < 2{,}80 + 1{,}96 \dfrac{1{,}10}{\sqrt{401}}$;

portanto, $2{,}69 < \mu < 2{,}91$.

(c) $\bar{x}$ (não fumantes) $= 0{,}54$, $\bar{x}$ (cachimbo ou charuto) $= 1{,}02$.

31. (a) $0{,}96 - 1{,}96 \sqrt{\dfrac{(0{,}96)\ (0{,}04)}{481}} < p < 0{,}96 +$

$+ 1{,}96 \sqrt{\dfrac{(0{,}96)\ (0{,}04)}{481}}$; portanto, $0{,}94 < p < 0{,}98$.

(b) $0{,}95 - 1{,}96 \sqrt{\dfrac{(0{,}95)\ (0{,}05)}{95}} < p < 0{,}95 +$

$+ 1{,}96 \sqrt{\dfrac{(0{,}95)\ (0{,}05)}{95}}$; portanto, $0{,}91 < p < 0{,}99$.

(c) Sim.

CAPÍTULO 8

1. Supondo que o indivíduo seja inocente até prova em contrário, a hipótese a ser testada é que o indivíduo é inocente; portanto, o erro do tipo I é condenar um inocente e o do tipo II é deixar livre o culpado. A sociedade considera o erro do tipo I muito mais sério que o do tipo II.

2. Considere-se que o teste de uma cura proposta de câncer seja inútil. O erro do tipo II é o de se decidir ser inútil, quando, de fato, é uma cura; esse é um erro muito sério. Logo se descobrirá uma afirmação falsa, se um erro do tipo I for cometido.

3. $\alpha = \dfrac{1}{4}$, $\beta = \dfrac{3}{4}$.

4.

x	0	1	2
$P_0\{x\}$	0,25	0,50	0,25
$P_1\{x\}$	0,09	0,42	0,49

$\alpha = P_0\{x = 2\} = 0{,}25$.

$\beta = P_1\{x \neq 2\} = 0{,}51$.

5. $\alpha = P_0\{x = 0\} = 0{,}25$, $\beta = P_1\{x \neq 0\} = 0{,}91$; portanto, $x = 2$ é uma região crítica muito melhor do que $x = 0$.

6. $z = \dfrac{82 - 86}{1{,}6} = -2{,}50$; portanto, rejeite H_0.

7. $z = \dfrac{82 - 86}{3{,}2} = -1{,}25$; portanto, aceite H_0.

8. $z = \dfrac{395 - 400}{2} = -2{,}50$; portanto, rejeite H_0 em favor de

$H_1 : \mu < 400$.

9. $z = \dfrac{68 - 64}{8/\sqrt{55}} = \dfrac{4}{1{,}08} = 3{,}7$; portanto, rejeite H_0 em favor de $H_1 : \mu > 64$.

10. $z = \dfrac{7 - 8}{1{,}4/\sqrt{50}} = \dfrac{-1}{0{,}20} = -5$; portanto, rejeite H_0. Não justificado.

11. $\bar{x} = 71{,}6$, $s = 31{,}4$; os limites são, portanto, 30 — 114. Parece existir controle, mas os dados são poucos demais para podermos tirar conclusões fidedignas.

13. $\bar{x} = 0{,}26$, $s = 0{,}167$, $z = \dfrac{0{,}26 - 0}{0{,}167 / \sqrt{100}} = 15{,}6$; portanto, rejeite $H_0 : \mu = 0$. Há tendenciosidade em favor do produto que é experimentado em primeiro lugar.

14. $z = \dfrac{52 - 40}{\sqrt{240 \cdot \frac{1}{6} \cdot \frac{5}{6}}} = 2{,}08$; portanto, rejeite $H_0 : p = \frac{1}{6}$. Sim.

15. $z = \dfrac{56 - 48}{\sqrt{120\ (0{,}4)\ (0{,}6)}} = 1{,}49$; portanto, aceite H_0. Não.

16. $z = \dfrac{120 - 100}{\sqrt{200 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = 2{,}83$; portanto, rejeite H_0. A mistura é forte demais.

17. Honesto em sentido matemático significa que $p = \frac{1}{2}$, ao passo que honesto, do ponto de vista prático, significa que p é aproximadamente igual a $\frac{1}{2}$.

18. Os lançamentos podem não ser aleatórios na natureza. É muito difícil eliminarmos os erros humanos em seqüências de tentativas.

19. $\hat{p} = 0{,}0255$; portanto, os limites são 0,0105 e 0,0405. Usando percentagens, multiplique estes limites por 100. Fora de controle nos dias numerados com 18, 22 e 38.

20. $z = \dfrac{20 - 22}{\sqrt{\frac{64}{100} + \frac{49}{100}}} = -1{,}88$; portanto, aceite H_0.

21. $z = \dfrac{73{,}4 - 70{,}2}{\sqrt{\frac{81}{50} + \frac{100}{50}}} = 1{,}68$; portanto, aceite H_0.

22. $z = \dfrac{54 - 58}{\sqrt{\frac{36}{40} + \frac{81}{50}}} = -2{,}52$; portanto, rejeite H_0.

23. $z = \dfrac{15{,}2 - 14{,}9}{\sqrt{\frac{(2{,}1)^2}{50} + \frac{(2{,}0)^2}{50}}} = -0{,}73$; portanto, aceite H_0.

24. $z = \dfrac{1{,}38 - 1{,}48}{\sqrt{\frac{(1{,}23)^2}{407} + \frac{(1{,}14)^2}{411}}} = -1{,}21$; portanto, aceite H_0. Nenhum efeito.

25. $z = \dfrac{0{,}22 - 0{,}17}{\sqrt{(0{,}19)\ (0{,}81)\left[\frac{1}{200} + \frac{1}{300}\right]}} = 1{,}40$; portanto, aceite H_0. Não.

26. $z = \dfrac{0{,}28 - 0{,}25}{\sqrt{(0{,}26)\ (0{,}74)\left[\frac{1}{200} + \frac{1}{300}\right]}} = 0{,}75$; portanto, aceite H_0. Não.

27. $z = \dfrac{0{,}14 - 0{,}08}{\sqrt{(0{,}11)\ (0{,}89)\left[\frac{1}{450} + \frac{1}{500}\right]}} = 2{,}95$; portanto, rejeite H_0.

28. $z = \dfrac{0{,}25 - 0{,}16 - d}{\sqrt{\frac{(0{,}25)\ (0{,}75)}{200} \quad \frac{(0{,}16)\ (0{,}84)}{400}}} = \dfrac{0{,}09 - d}{0{,}036} = 1{,}64$; portanto, $d = 0{,}03$.

30. (a) $z = \dfrac{0{,}76 - 0{,}61}{\sqrt{(0{,}68)\ (0{,}32)\ [\frac{1}{771} + \frac{1}{771}]}} = 6{,}3$; portanto, rejeite H_0.

(b) $z = \dfrac{0{,}79 - 0{,}75}{\sqrt{(0{,}77)\ (0{,}23)\ [\frac{1}{514} + \frac{1}{514}]}} = 1{,}52$; portanto, aceite H_0.

31. $t = \dfrac{42 - 44}{6}\sqrt{20} = -1{,}49$; portanto, aceite H_0.

32. $\bar{x} = 9{,}77$, $s = 1{,}44$, $t = \dfrac{9{,}77 - 9}{1{,}44}\sqrt{15} = 2{,}07$. Uma vez que $t_0 = 1{,}76$ para um teste unilateral, rejeite H_0.

33. $t = \dfrac{20 - 24}{\sqrt{9\,(7)^2 + 11\,(6)^2}}\sqrt{\dfrac{10.12.20}{22}} = -1{,}44$; portanto, aceite H_0.

34. $\bar{x}_1 = 57{,}9$, $\bar{x}_2 = 55{,}8$, $s_1 = 5{,}59$, $s_2 = 4{,}59$,

$t = \dfrac{57{,}9 - 55{,}8}{\sqrt{9(5{,}59)^2 + 9(4{,}59)^2}}\sqrt{\dfrac{10.10.18}{20}} = 0{,}92$; portanto, aceite H_0.

35. $\bar{x}_1 = 56{,}3$, $\bar{x}_2 = 62{,}3$, $s_1 = 9{,}34$, $s_2 = 10{,}16$,

$t = \dfrac{56{,}3 - 62{,}3}{\sqrt{11(9{,}34)^2 + 11(10{,}16)^2}}\sqrt{\dfrac{12.12.22}{24}} = -1{,}51$; portanto, aceite H_0.

36. $\bar{x} = 7{,}5$, $s = 11{,}3$, $t = \dfrac{7{,}5}{11{,}3}\sqrt{12} = 2{,}30$. Como $t_0 = 1{,}796$ para um teste unilateral, rejeite H_0.

37. (a) $z = \dfrac{7{,}04 - 7{,}19}{\sqrt{\frac{(1{,}2)^2}{158} + \frac{(0{,}9)^2}{94}}} = -1{,}13$; portanto, aceite H_0.

(b) $z = \dfrac{23{,}3 - 21{,}9}{\sqrt{\frac{(2{,}8)^2}{158} \quad \frac{(3{,}0)^2}{94}}} = 3{,}67$; portanto, rejeite H_0.

(c) uma característica de futura obesidade é um grande aumento de peso durante o primeiro ano de vida.

38. $z = \dfrac{0{,}08 - 0{,}46}{\sqrt{(0{,}27)\ (0{,}73)\ [\frac{1}{38} + \frac{1}{39}]}} = -3{,}76$; portanto, rejeite H_0.

39. $z = \dfrac{0{,}46 - 0{,}11}{\sqrt{(0{,}19)\ (0{,}81)\ [\frac{1}{13} + \frac{1}{45}]}} = 2{,}83$; portanto, rejeite H_0.

CAPÍTULO 9

1. (c) $\Sigma x = 828$, $\Sigma y = 1\,882$, $\Sigma x^2 = 57\,266$; $\Sigma y^2 = 302\,010$, $\Sigma xy = 130\,753$; portanto, $\bar{x} = 69$, $\bar{y} = 156{,}833$, $s_x = 3{,}490$, $s_y = 24{,}954$, e

$$r = \frac{130\,753 - 12(69)\ (156{,}833)}{11(3{,}49)\ (24{,}954)} = 0{,}93.$$

2. $\Sigma x = 3\,697$, $\Sigma y = 41$, $\Sigma x^2 = 722\,611$, $\Sigma y^2 = 97{,}72$, $\Sigma xy = 7\,930$; portanto, $\bar{x} = 184{,}85$, $\bar{y} = 2{,}05$, $s_x = 45{,}434$, $s_y = 0{,}848$, e

$$r = \frac{7\,930 - 20(184{,}85)\ (2{,}05)}{19(45{,}434)\ (0{,}848)} = 0{,}48.$$

3. A fórmula mostra que o intercâmbio entre x e y não tem efeito sobre o valor de r.

4. (a) 0,9, (b) 0,3, (c) −0,7.

5. (a) 0,4, (b) −0,6, (c) −0,5.

6. O número de acidentes anuais de motoristas que já tiveram acidentes decresce com a idade.

7. Se x expressar maré alta e y expressar densidade de tráfego tanto x como y tenderão a ser grandes em torno de 8:00 A.M. e pequeno em torno de 2:00 P.M.

8. $\Sigma x = 1\,897$, $\Sigma y = 18{,}6$, $\Sigma x^2 = 364\,973$, $\Sigma y^2 = 40{,}6$, $\Sigma xy = 3\,501{,}2$; portanto $\bar{x} = 189{,}7$, $\bar{y} = 1{,}86$, $s_x = 23{,}83$, $s_y = 0{,}817$, e

$$r = \frac{3\,501{,}2 - 10\,(189{,}7)\,(1{,}86)}{9(23{,}83)\,(0{,}817)} = -\,0{,}16.$$

Eliminando os indivíduos com quociente de inteligência baixa e alta, remove-se a inteligência como fator que afeta as médias.

10. Para n= 30, $r_0 = 0{,}361$; portanto, aceita-se H_0.

11. Dever-se-ia usar aqui um teste unilateral; portanto n = 18.

12. Para n = 35, $r_0 = 0{,}283$; portanto, rejeita-se H_0.

13. Para n = 30, $|r_0| = 0{,}361$ para um teste bilateral, enquanto $|r_0| = 0{,}306$ para um teste unilateral; portanto, aceita-se no primeiro caso e rejeita-se no segundo.

16. (a) y = \$ 700, \$ 1.300 (b) A fórmula provavelmente foi derivada para famílias médias com rendas regulares e, portanto, é improvável que sejam aplicáveis para este valor de x.

17. $\Sigma x = 15$, $\Sigma y = 25$, $\Sigma x^2 = 55$, $\Sigma xy = 86$, portanto, $\bar{y} = 5$ e

$$b = \frac{86 - 5\,(3)\,(5)}{55 - 5(3)^2} = 1{,}1.$$

Portanto, $y' = 5 + 1{,}1\,(x - 3)$ ou $y' = 1{,}7 + 1{,}1x$.

18. $\Sigma x = 803$, $\Sigma y = 3\,906$, $\Sigma x^2 = 55\,069$, $\Sigma xy = 264\,385$; portanto, $\bar{x} = 66{,}92$, $\bar{y} = 325{,}50$, e

$$b = \frac{264\,385 - 12(66{,}92)\,(325{,}5)}{55\,069 - 12(66{,}92)^2} = 2{,}25.$$

Portanto, $y' = 325{,}5 + 2{,}25\,(x - 66{,}92)$ ou
$y' = 174{,}9 + 2{,}25x$.

19. $\Sigma x = 3\,616$, $\Sigma y = 4\,404$, $\Sigma x^2 = 1\,191{,}462$, $\Sigma xy = 1\,339\,478$; portanto, $\bar{x} = 301{,}33$, $\bar{y} = 367$, e

$$b = \frac{1\,339\,478 - 12(301{,}33)\,(367)}{1\,191\,462 - 12(301{,}33)^2} = 0{,}122.$$

Portanto, $y' = 367 + 0{,}122(x - 301{,}33)$, ou
$y' = 330{,}2 + 0{,}122x$.

24.

y	2	5	4	8	6
y'	2,8	3,9	5,0	6,1	7,2
y − y'	−0,8	1,1	−1,0	1,9	−1,2
(y − y')²	0,64	1,21	1,00	3,61	1,44

$$s_e = \sqrt{\frac{7{,}90}{3}} = 1{,}62.$$

25.

y	293	349	368	301	340	308	354	313
y'	294	332	364	299	350	319	335	294
y − y'	−1	17	4	2	−10	−11	19	19

y	322	334	377	247
y'	359	326	332	301
y − y'	−37	8	45	−54

$s_e = \sqrt{\frac{7627}{10}} = 27{,}6$; portanto, 25% excede s_e e 0% excede $2\,s_e$

26. Ambos estão baseados em pequenas amostras; portanto, as estimativas não são fidedignas.

27. $2 - 2{,}45\,\frac{0{,}2}{\sqrt{50}} < \beta < 2 + 2{,}45\,\frac{0{,}2}{\sqrt{50}}$; portanto,

$1{,}93 < \beta < 2{,}07$.

28. $2{,}25 - 2{,}23\,\frac{27{,}6}{\sqrt{1\,329{,}6}} < \beta < 2{,}25 + 2{,}23\,\frac{27{,}6}{\sqrt{1\,329{,}6}}$; portanto,

$0{,}56 < \beta < 3{,}94$.

29. $t = \frac{1{,}1 - 1{,}0}{1{,}62}\sqrt{10} = 0{,}20$; portanto, aceita-se H_0.

30. $t = \frac{2{,}25 - 2{,}5}{27{,}6}\sqrt{1\,329{,}6} = -\,0{,}33$; portanto, aceita-se H_0.

31. Escolha os "x" tão afastados quanto possível de modo a fazer $\Sigma(x_i - \bar{x})^2$ tão grande quanto possível; portanto, escolha cinco valores em $x = 0$ e cinco valores em $x = 10$.

32. $\Sigma x = 59$, $\Sigma y = 542$, $\Sigma x^2 = 371$, $\Sigma y^2 = 28\,820$; $\Sigma xy = 3\,197$; portanto, $\bar{x} = 4{,}92$, $\bar{y} = 45{,}17$, $s_x = 2{,}71$, $s_y = 19{,}85$.

(b) $$r = \frac{3\,197 - 12(4{,}92)\,(45{,}17)}{11(2{,}71)\,(19{,}85)} = 0{,}90.$$

(c) $r_0 = 0{,}576$; portanto, rejeita-se H_0.

(d) $$b = \frac{3\,197 - 12(4{,}92)\,(45{,}17)}{371 - 12(4{,}92)^2} = 6{,}58;$$ portanto,

$$y' = 45{,}17 + 6{,}58\,(x - 4{,}92) = 12{,}8 + 6{,}58x.$$

(e)

y	15	75	55	42	33	45	55	17	32	80	48	45
y'	19	72	52	39	46	59	59	19	26	65	52	33
$y - y'$	−4	3	3	3	−13	−14	−4	−2	6	15	−4	12

(f) $$\sqrt{\frac{849}{10}} = 9{,}2.$$

(g) 0%

33. $\Sigma x = 143$, $\Sigma x = 156{,}6$, $\Sigma x^2 = 1\,755$, $\Sigma xy = 1\,826{,}8$; portanto, $\bar{x} = 11$, $\bar{y} = 12{,}05$ e

(b) $$b = \frac{1\,826{,}8 - 13(11)\,(12{,}05)}{1\,755 - 13(11)^2} = 0{,}570;$$

Portanto, $y' = 12{,}05 + 0{,}570\,(x - 11) = 5{,}78 + 0{,}570x$.

(d) $y' = 15{,}2$.

(e)

y	7,6	9,5	9,3	10,3	11,1	12,1	13,3	12,7	13,0	13,8	14,6	14,6	14,7
y'	8,6	9,2	9,8	10,3	10,9	11,5	12,0	12,6	13,2	13,8	14,3	14,9	15,5
$y - y'$	−1,0	0,3	−0,5	0,0	0,2	0,6	1,3	0,1	−0,2	0,0	0,3	−0,3	−0,8

(f) $$s_e = \sqrt{\frac{4{,}30}{11}} = 0{,}63.$$

(g) Há 4 pontos que caem fora da faixa, que é mais ou menos o que se espera.

(h) $$t = \frac{0{,}57 - 0{,}50}{0{,}63}\sqrt{182} = 1{,}50;$$ portanto, aceite H_0.

(i) $$0{,}57 - 1{,}796\frac{0{,}63}{\sqrt{182}} < \beta < 0{,}57 + 1{,}796\frac{0{,}63}{\sqrt{182}};$$ portanto,

$$0{,}49 < \beta < 0{,}65.$$

34. $\Sigma x = 2\,137$, $\Sigma y = 2\,170$, $\Sigma x^2 = 156\,513$, $\Sigma y^2 = 161\,146$, $\Sigma xy = 157\,708$; portanto, $\bar{x} = 71{,}233$, $\bar{y} = 72{,}333$, $s_x = 12{,}159$, $s_y = 12{,}010$.

(b) $$r = \frac{157\,708 - 30(71{,}233)\,(72{,}333)}{29(12{,}159)\,(12{,}010)} = 0{,}74.$$

(c) $r_0 = 0{,}361$; portanto, rejeite H_0.

(d) $$b = \frac{157\,708 - 30(71{,}233)\,(72{,}333)}{156\,513 - 30(71{,}233)^2} = 0{,}731;$$ portanto,

$$y' = 72{,}333 + 0{,}731\,(x - 71{,}233) = 20{,}26 + 0{,}731x.$$

(e) $$s_e = \sqrt{\frac{161\,146 - 20{,}26\,(2\,170) - 0{,}731\,(157\,700)}{28}} = 8{,}2.$$

CAPÍTULO 10

1. $n = 144$, $p_i = \frac{9}{16}, \frac{3}{16}, \frac{4}{16}$; portanto,

o_i	74	32	38
e_i	81	27	36

$$\chi^2 = \frac{49}{81} + \frac{25}{27} + \frac{4}{36} = 1{,}6.$$

Como $\chi^2_0 = 5{,}99$, a teoria é substanciada.

2. $n = 770$, $p_i = 0{,}16,\ 0{,}48,\ 0{,}20,\ 0{,}16$; portanto,

o_i	180	360	130	100
e_i	123	370	154	123

$$\chi^2 = \frac{(57)^2}{123} + \frac{(10)^2}{370} + \frac{(24)^2}{154} + \frac{(23)^2}{123} \cong 35.$$

$\chi^2_0 = 7{,}8$; portanto, a distribuição dos tipos sanguíneos é diferente.

3. $n = 100$, $p_i = 0{,}75,\ 0{,}20,\ 0{,}05$; portanto,

o_i	65	22	13
e_i	75	20	5

$$\chi^2 = \frac{100}{75} + \frac{4}{20} + \frac{64}{5} = 14{,}3.$$

$\chi^2_0 = 5{,}99$; portanto, as proporções postuladas são incorretas.

4.

o_i	65	35
e_i	75	25

$$\chi^2 = \frac{100}{75} + \frac{100}{25} = 5{,}3.$$

$\chi^2_0 = 3{,}8$; portanto, rejeite H_0. Embora H_0 seja rejeitado aqui também, a rejeição é menos decisiva porque a grande diferença que ocorre na terceira célula está mascarada pelas freqüências quase iguais na segunda célula.

5.

o_i	22	28	17	18	7	8
e_i	25	25	20	15	10	5

$$\chi^2 = \frac{9}{25} + \frac{9}{25} + \frac{9}{20} + \frac{9}{15} + \frac{9}{10} + \frac{9}{5} = 4{,}5.$$

$\chi^2_0 = 11{,}1$; portanto, a amostra é compatível.

6.

o_i	11	8	17	5	13	9	17	6	10	4
e_i	10	10	10	10	10	10	10	10	10	10

$$\chi^2 = \frac{1}{10}[1 + 4 + 49 + 25 + 9 + 1 + 49 + 16 + 0 + 36] = 19.$$

$\chi^2_0 = 16{,}9$; portanto, rejeite a hipótese de que a taxa de acidentes era constante ao longo deste período de tempo.

8.

o_i	19	10	6	5	2	0
e_i	12	10	8	6	4	2

. Combine as duas últimas células.

Então, $\chi^2 = \frac{49}{12} + \frac{4}{8} + \frac{1}{6} + \frac{16}{6} = 7{,}4$. Como $\chi^2_0 = 9{,}49$, aceita-se a hipótese de que os pagamentos são corretos.

9.

o_i	29	13
e_i	22	20

$$\chi^2 = \frac{49}{22} + \frac{49}{20} = 4{,}7.$$

Como $\chi^2_0 = 3{,}84$, a crença de que os pagamentos são baixos demais para buracos difíceis é justificada.

10. $P\{x\} = \frac{4!}{x!(4-x)!}\left(\frac{1}{2}\right)^4$ dá

x	0	1	2	3	4
P{x}	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

;

Portanto,

o_i	4	10	8	7	3
e_i	2	8	12	8	2

. Combine as primeiras duas células e as duas últimas para obter

o_i	14	8	10
e_i	10	12	10

$$\chi^2 = \frac{16}{10} + \frac{16}{12} = 2{,}9.$$

Como $\chi^2_0 = 5{,}99$, aceite o modelo binomial.

11.

20 (12)	10 (12)	10 (16)	40
10 (18)	20 (18)	30 (24)	60
30	30	40	100

$$\chi^2 = \frac{64}{12} + \frac{4}{12} + \frac{36}{16} + \frac{64}{18} + \frac{4}{18} + \frac{36}{24} = 13{,}2.$$

$\chi^2_0 = 5{,}99$; portanto, rejeite a independência.

12.

50 (47)	10 (11)	22 (24)	82
44 (47)	12 (11)	26 (24)	82
94	22	48	164

$$\chi^2 = \frac{9}{47} + \frac{1}{11} + \frac{4}{24} + \frac{9}{47} + \frac{1}{11} + \frac{4}{24} = 0{,}90.$$

$\chi^2_0 = 5{,}99$; portanto, aceite a hipótese de que a droga não é melhor que a pílula de açúcar para curar resfriados.

13.

62 (60)	34 (36)	96
84 (81)	44 (47)	128
24 (29)	22 (17)	46
170	100	270

$$\chi^2 = \frac{4}{60} + \frac{4}{36} + \frac{9}{81} + \frac{9}{47} + \frac{25}{29} + \frac{25}{17} = 2{,}8.$$

$\chi^2_0 = 5{,}99$; portanto, aceite a hipótese de que eles têm a mesma opinião.

14.

	A	N	
T	166 (151)	353 (368)	519
N	104 (119)	304 (289)	408
	270	657	927

$$\chi^2 = \frac{225}{151} + \frac{225}{368} + \frac{225}{119} + \frac{225}{289} = 4{,}8.$$

$\chi^2_0 = 3{,}84$; portanto, rejeite a hipótese da não influência. A comparação das freqüências observadas e esperadas mostra que o tratamento foi prejudicial e não benéfico.

15. $\hat{p}_1 = \frac{166}{519} = 0{,}320$, $\hat{p}_2 = \frac{104}{408} = 0{,}255$, $\hat{p} = \frac{270}{927} = 0{,}29$,

$$z = \frac{0{,}320 - 0{,}255}{\sqrt{(0{,}29)\ (0{,}71)\ [\frac{1}{519} + \frac{1}{408}]}} = 2{,}2;$$

portanto, rejeite a hipótese da não diferença nas proporções de efeitos posteriores para os dois grupos. Pode-se demonstrar que $z^2 = \chi^2$ para este tipo de problema.

16.

18 (13)	17 (19)	5 (8)	40
26 (27)	38 (37)	16 (16)	80
6 (10)	15 (14)	9 (6)	30
50	70	30	150

$$\chi^2 = \frac{25}{13} + \frac{4}{19} + \frac{9}{8} + \frac{1}{27} + \frac{1}{37} + \frac{16}{10} + \frac{1}{14} + \frac{9}{6} = 6{,}5.$$

$\chi_0^2 = 9{,}5$; portanto, aceite a hipótese de nenhuma relação entre estas duas variáveis.

17. (a) $p_i : \frac{267}{711}, \frac{83}{711}, \frac{82}{711}, \frac{80}{711}, \frac{199}{711}$; portanto, as freqüências esperadas são obtidas multiplicando-se estas probabilidades por 76. Isto nos dá

o_i	22	7	8	6	33
e_i	29	9	9	9	21

$$\chi^2 = \frac{49}{29} + \frac{4}{9} + \frac{1}{9} + \frac{9}{9} + \frac{144}{21} = 10{,}1.$$

$\chi_0^2 = 9{,}5$; portanto, rejeite a hipótese de que a taxa de acidente é a mesma para os dois grupos.

(b)

o_i	22	21
e_i	22	21

$\chi^2 = 0$; portanto, aceite a hipótese de que as taxas de acidentes são as mesmas para os dois grupos.

(c) Os indivíduos mais velhos que não são nem senis nem possuidores de distúrbios cardiovasculares têm tanta segurança quanto os que estão na faixa dos 30-59 anos.

18.

o_i	80	94	89	84	91	81	64	123
e_i	88	88	88	88	88	88	88	88

(a) $\chi^2 = \frac{1}{88}[64 + 36 + 1 + 16 + 9 + 49 + 576 + 1\,225] = 21{,}5.$

$\chi_0^2 = 14{,}1$; portanto, rejeite a hipótese de uma taxa de admissão constante.

(b)

o_i	583	123
e_i	618	88

$$\chi^2 = \frac{1\,225}{618} + \frac{1\,225}{88} = 15{,}9.$$

$\chi_0^2 = 3{,}84$; portanto, rejeite a hipótese da mesma taxa.

19. Combine as duas primeiras e as duas últimas linhas para obter

28 (10)	26 (33)	5 (16)	59
19 (35)	130 (113)	56 (57)	205
17 (18)	50 (58)	38 (29)	105
4 (5)	15 (18)	13 (9)	32
68	221	112	401

$$\chi^2 = \frac{324}{10} + \frac{36}{33} + \frac{121}{16} + \frac{256}{35} + \frac{289}{113} + \frac{1}{57} + \frac{1}{18} + \frac{64}{58} + \frac{81}{29} + \frac{1}{5} + \frac{9}{18} + \frac{16}{9} = 57.$$

$\chi_0^2 = 12{,}6$; portanto, rejeite a hipótese de nenhuma relação entre o grau de enfisema e a quantidade de fumo.

CAPÍTULO 11

1. $\bar{x}_{.1} = 34$, $\bar{x}_{.2} = 36$, $\bar{x}_{.3} = 37{,}67$, $\bar{x}_{.4} = 33{,}17$, $\bar{x} = 35{,}21$,
$\Sigma(\bar{x}_{.j} - \bar{x})^2 = 12{,}3$, $\Sigma\Sigma(x_{ij} - \bar{x}_{.j})^2 = 36{,}1$,

$$F = \frac{6.4.5}{3} \frac{12{,}3}{36{,}1} = 13{,}6.$$

Como $\nu_1 = 3$, $\nu_2 = 20$ e $F_0 = 3{,}1$, rejeite a hipótese de que os catalisadores não têm efeito.

2. $\bar{x}_{.1} = 7{,}45$, $\bar{x}_{.2} = 6{,}75$, $\bar{x}_{.3} = 7{,}20$, $\bar{x}_{.4} = 4{,}75$, $\bar{x} = 6{,}54$,
$\Sigma(\bar{x}_{.j} - \bar{x})^2 = 4{,}51$, $\Sigma\Sigma(x_{ij} - \bar{x}_{.j})^2 = 128{,}4$.

$$F = \frac{4.4.3}{3} \frac{4{,}51}{128{,}4} = 0{,}56.$$

Como $\nu_1 = 3$, $\nu_2 = 12$ e $F_0 = 3{,}5$, aceite a hipótese de nenhuma diferença devido aos métodos de pulverização.

3. $\bar{x}_{.1} = 47$, $\bar{x}_{.2} = 44$, $\bar{x}_{.3} = 50$, $\bar{x} = 47$,
$\Sigma(\bar{x}_{.j} - \bar{x})^2 = 18$, $\Sigma\Sigma(x_{ij} - \bar{x}_{.j})^2 = 1\,888$,

$$F = \frac{8.3.7}{2} \frac{18}{1\,888} = 0{,}80$$

Como $\nu_1 = 2$, $\nu_2 = 21$ e $F_0 = 3{,}47$, aceite a hipótese de que a exatidão é independente do tipo de visão (disposição dos olhos).

4. $\bar{x}_1 = 7{,}133$, $\bar{x}_2 = 4{,}75$, $\bar{x} = 6{,}54$,

$s_1^2 = 10{,}88$, $s_2^2 = 3{,}23$,

$$t = \frac{7{,}133 - 4{,}75}{\sqrt{11(10{,}88) + 3(3{,}23)}} \sqrt{\frac{12 \cdot 4 \cdot 14}{16}} = 1{,}36;$$

portanto, aceite a hipótese de "não diferença na produção" devido à pulverização

5. $\bar{x}_{\cdot 1} = 41$, $\bar{x}_{\cdot 2} = 38{,}8$, $\bar{x}_{\cdot 3} = 47{,}6$, $\bar{x}_{\cdot 4} = 36{,}6$, $\bar{x} = 41$,

$\bar{x}_1 = 41{,}25$, $\bar{x}_2 = 45{,}25$, $\bar{x}_3 = 36{,}5$, $\bar{x}_4 = 40{,}0$, $\bar{x}_5 = 42{,}0$,

$\Sigma (\bar{x}_{\cdot j} - \bar{x})^2 = 67{,}76$, $\Sigma (\bar{x}_i - \bar{x})^2 = 40{,}38$,

$\Sigma\Sigma (x_{ij} - \bar{x})^2 = 574$, $\Sigma\Sigma (x_{ij} - \bar{x}_i - \bar{x}_{\cdot j} + \bar{x})^2 = 574 - 4(40{,}38) - 5(67{,}76) = 73{,}68$

(a) $F = 5 \cdot 4 \dfrac{67{,}76}{73{,}68} = 18{,}4.$

Como $\nu_1 = 3$, $\nu_2 = 12$ e $F_0 = 3{,}5$, rejeite a hipótese de que os tipos não diferem.

(b) $F = 4 \cdot 3 \dfrac{40{,}38}{73{,}68} = 6{,}6.$

Como $\nu_1 = 4$, $\nu_2 = 12$ e $F_0 = 3{,}3$, rejeite a hipótese de que os trabalhadores são igualmente produtivos.

6. $\bar{x}_1 = 6{,}38$, $\bar{x}_2 = 4{,}78$, $\bar{x}_3 = 7{,}02$, $\bar{x}_4 = 7{,}98$, $\bar{x} = 6{,}54$,

$\Sigma (\bar{x}_i - \bar{x})^2 = 5{,}43$, $\Sigma\Sigma (x_{ij} - \bar{x})^2 = 145{,}9$,

$\Sigma\Sigma (x_{ij} - \bar{x}_i - \bar{x}_{\cdot j} + \bar{x})^2 = 145{,}9 - 4(5{,}43) - 4(4{,}51) = 106{,}1.$

$$F = 4 \cdot 3 \frac{4{,}51}{106{,}1} = 0{,}51.$$

Como $\nu_1 = 3$, $\nu_2 = 9$ e $F_0 = 3{,}9$, aceite a mesma conclusão que a do problema 2.

7. $F = 5 \cdot 4 \dfrac{1\,286}{2388} = 10{,}8.$

Como $\nu_1 = 3$, $\nu_2 = 12$ e $F_0 = 3{,}5$, rejeite a hipótese de falta de diferenças.

8. Como $\Sigma\Sigma (x_{ij} - \bar{x})^2 = \Sigma\Sigma (\bar{x}_{\cdot j} - \bar{x})^2 + \Sigma\Sigma (x_{ij} - \bar{x}_{\cdot j})^2$,

$\Sigma\Sigma (x_{ij} - \bar{x}_{\cdot j})^2 = 21\,530 - 4(3178) = 8\,818.$

$$F = \frac{4 \cdot 5 \cdot 3}{4} \frac{3\,178}{8\,818} = 5{,}4.$$

Como $\nu_1 = 4$, $\nu_2 = 15$ e $F_0 = 3{,}1$, rejeite a hipótese de falta de diferença entre os fertilizantes. A solução do texto na qual as diferenças são eliminadas é mais decisiva aqui.

9. $\bar{x}_{\cdot 1} = 54$, $\bar{x}_{\cdot 2} = 62$, $\bar{x}_{\cdot 3} = 58$, $\bar{x}_{\cdot 4} = 54$, $\bar{x} = 57$,

$\bar{x}_1 = 55$, $\bar{x}_2 = 61$, $\bar{x}_3 = 55$,

$\Sigma (\bar{x}_i - \bar{x})^2 = 24$, $\Sigma (\bar{x}_{\cdot j} - \bar{x})^2 = 44$, $\Sigma\Sigma (x_{ij} - \bar{x})^2 = 252$,

$\Sigma\Sigma (x_{ij} - \bar{x}_i - \bar{x}_{\cdot j} + \bar{x})^2 = 252 - 4(24) - 3(44) = 24.$

(a) $F = 3 \cdot 2 \dfrac{44}{24} = 11.$

Como $\nu_1 = 3$, $\nu_2 = 6$, e $F_0 = 4{,}8$, rejeite a hipótese de nenhum efeito catalítico.

(b) $F = 4 \cdot 3 \dfrac{24}{24} = 12.$ Como $\nu_1 = 2$, $\nu_2 = 6$, e $F_0 = 5{,}1$, rejeite a hipótese de nenhum efeito de temperatura.

CAPÍTULO 12

1. 9 diferenças positivas; portanto $x = 9$ e $n = 29$.

$$z = \frac{9{,}5 - 14{,}5}{\sqrt{29 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = -1{,}86.$$ Como $z_0 = -1{,}64$, rejeite a hipótese de que a idade média seja de 23 em favor de uma idade mais baixa.

2. 10 diferenças positivas; portanto, $x = 10$ e $n = 15$.

$$z = \frac{9{,}5 - 7{,}5}{\sqrt{15 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = 1{,}03$$

Como $z_0 = 1{,}96$, aceite a hipótese da não-diferença.

3. $\bar{x} = 1{,}40$, $s = 2{,}95$,

$$t = \frac{1{,}40}{2{,}95} \sqrt{15} = 1{,}84.$$

Como as duas casas de zero foram eliminadas aqui também, $\nu = 14$ e $t_0 = 2{,}145$; portanto, aceite a hipótese da não-diferença. O valor t está muito próximo do valor crítico quando comparado ao valor $z = 1{,}03$ do problema 2, indicando que o teste t é mais sensível ao deslocamento na média do que o teste de sinal. Uma amostra maior poderia revelar uma diferença real.

4. 9 diferenças positivas; portanto, x = 9 e n = 12.

$$z = \frac{8{,}5 - 6}{\sqrt{12 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = 1{,}44.$$ Para um teste unilateral $t_0 = 1{,}64$; portanto, aceite a hipótese da não-diferença. O teste t produziu o valor $t = 2{,}30$ e rejeitou-se H_0.

5. 21 diferenças positivas; portanto, x = 21 e 30.

$$z = \frac{20{,}5 - 15}{\sqrt{30 \cdot \frac{1}{2} \cdot \frac{1}{2}}} = 2{,}01.$$ Para um teste unilateral $z_0 = 1{,}64$; portanto, rejeita-se H_0.

6. (a) R = 64. A região crítica para $\alpha = 0{,}052$ compõe-se de $R \leqslant 79$ e $R \geqslant 131$; portanto, rejeita-se H_0.

(b) $\mu_R = 105$, $\sigma_R = 13{,}2$, $z = \frac{64 - 105}{13{,}2} = -3{,}1$; portanto, rejeita-se H_0.

7. Os pares de valores não são valores que se combinam aqui; portanto, não se deve considerar estas diferenças particulares.

8. $R = 156$, $\mu_R = 168$, $\sigma_R = 20{,}5$,

$$z = \frac{156 - 168}{20{,}5} = 0{,}59;$$ portanto, aceite a hipótese da não-diferença.

9. $t = 4{,}10$, $t_0 = 2{,}10$; portanto, rejeite H_0. O teste t parece rejeitar H_0 mais facilmente, aqui, que o teste da soma dos postos.

10 $R = 122{,}5$, $\mu_R = 105$, $\sigma_R = 13{,}23$,

$$z = \frac{122{,}5 - 105}{13{,}23} = 1{,}32;$$ portanto, aceite H_0.

11. $R = 153{,}5$, $\mu_R = 168$, $\sigma_R = 20{,}5$,

$$z = \frac{153{,}5 - 168}{20.5} = -0{,}71;$$ portanto, aceite a hipótese da não-diferença nos níveis de tolerância.

12. $\Sigma (x_i - y_i)^2 = 192$, $r = 1 - 6 \frac{192}{12 \cdot 143} = 0{,}33$. Como $r_0 = 0{,}504$, aceite H_0: $\rho = 0$. Isto implica que os juízes não concordem sobre a qualidade da geléia.

13. $\Sigma (x_i - y_i)^2 = 84$, $r = 1 - 6 \frac{84}{17 \cdot 288} = 0{,}90$; portanto, rejeite H_0: $\rho = 0$. Os escores foram afetados pelo emparelhamento.

14. $r = 0{,}377$ se unilateral, $r \cong 0{,}48$ se bilateral. Este valor foi obtido por interpolação entre os valores de 0,05 e 0,01.

CAPÍTULO 13

1. $\Sigma x_1 = 45$, $\Sigma x_2 = 96$, $\Sigma y = 140$, $\Sigma x_1 x_2 = 506$, $\Sigma x_1 y = 825$, $\Sigma x_2 y = 1{,}593$, $\Sigma x_1^2 = 285$, $\Sigma x_2^2 = 1\,092$; portanto, as equações dos mínimos quadrados são:

$$10a_0 + 45a_1 + 96a_2 = 140$$
$$45a_0 + 285a_1 + 506a_2 = 825$$
$$96a_0 + 506a_1 + 1\,092a_2 = 1\,593.$$

Como soluções destas equações temos $a_0 = -0{,}600$, $a_1 = 1{,}725$, e $a_2 = 0{,}712$; portanto, a equação da regressão linear é

$$y = -0{,}60 + 1{,}725x_1 + 0{,}712x_2.$$

2. $\Sigma x_1 = 48$, $\Sigma x_2 = 66$, $\Sigma y = 192$, $\Sigma x_1 x_2 = 280$, $\Sigma x_1 y = 929$, $\Sigma x_2 y = 1\,137$, $\Sigma x_1^2 = 290$, $\Sigma x_2^2 = 428$; portanto, as equações de mínimos quadrados são:

$$12a_0 + 48a_1 + 66a_2 = 192$$
$$48a_0 + 290a_1 + 280a_2 = 929$$
$$66a_0 + 280a_1 + 428a_2 = 1\,137$$

As soluções destas equações são $a_0 = 5{,}18$, $a_1 = 1{,}500$ e $a_2 = 0{,}877$; portanto, a equação da função da regressão linear é

$$y = 5{,}18 + 1{,}500x_1 + 0{,}877x_2.$$

3. (a) Os dados do problema 1 necessários aqui são:

$$\Sigma x_1 = 45,\ \Sigma y = 140,\ \Sigma x_1^2 = 285,\ \Sigma x_1 y = 825.$$

As equações de mínimos quadrados a serem resolvidas são:

$$10a_0 + 45a_1 = 140$$
$$45a_0 + 285a_1 = 825.$$

A solução destas equações é $a_0 = 3{,}364$ e $a_1 = 2{,}364$; portanto, a função de regressão aqui é

$$y = 3{,}364 + 2{,}364x_1.$$

(b) Para esta função de regressão, os cálculos dão

y	0	4	9	18	10	16	20	20	18	25
y′	3,4	5,7	8,1	10,5	12,8	15,2	17,5	19,9	22,3	24,6
y − y′	−3,4	−1,7	0,9	7,5	−2,8	0,8	2,5	0,1	−4,3	0,4

$\Sigma\,(y - y')^2 \cong 105.$

Para a função de regressão do problema 1, os cálculos dão.

y	0	4	9	18	10	16	20	20	18	25
y′	0,8	8,2	7,1	13,1	9,1	15,9	20,4	21,4	19,6	24,2
y − y′	−0,8	−4,2	1,9	4,9	0,9	0,1	−0,4	−1,4	−1,6	0,8

$\Sigma\,(y - y')^2 \cong 52$. Como a soma dos quadrados é cerca da metade do tamanho quando x_2 é ignorado, x_2 é certamente benéfico.

4. $\Sigma x = 15$, $\Sigma x^2 = 55$, $\Sigma x^3 = 225$, $\Sigma x^4 = 979$, $\Sigma y = 170$, $\Sigma xy = 354$, $\Sigma x^2y = 1\,222$. As equações de mínimos quadrados são:

$$\begin{aligned} 6a_0 + 15a_1 + 55a_2 &= 170 \\ 15a_0 + 55a_1 + 225a_2 &= 354 \\ 55a_0 + 225a_1 + 979a_2 &= 1\,222. \end{aligned}$$

A solução destas equações é $a_0 = 40{,}1$, $a_1 = -6{,}55$, $a_2 = 0{,}50$; portanto, a equação da parábola é

$$y = 40{,}1 - 6{,}55x + 0{,}50x^2.$$

A parábola terá um mínimo entre 6 e 7, após o que y aumentará. É irrazoável esperar que o número de homens-hora aumente com o aumento da experiência.

5. $\Sigma x = 143$, $\Sigma x^2 = 1\,755$, $\Sigma x^3 = 23{,}309$, $\Sigma x^4 = 327{,}015$, $\Sigma y = 156{,}6$, $\Sigma xy = 1\,826{,}8$, $\Sigma x^2y = 233{,}668$.

(b) $\bar{x} = 11$, $\bar{y} = 12{,}05$,

$$b = \frac{1\,826{,}8 - 143(12{,}05)}{1\,755 - 143(11)} = 0{,}570.$$

$$y = 12{,}05 + 0{,}570(x - 11) = 5{,}78 + 0{,}570x.$$

(c) As equações de mínimos quadrados são

$$\begin{aligned} 13a_0 + 143a_1 + 1\,755a_2 &= 156{,}6 \\ 143a_0 + 1\,755a_1 + 23\,309a_2 &= 1\,826{,}8 \\ 1\,755a_0 + 23\,309a_1 + 327\,015a_2 &= 23\,366{,}8. \end{aligned}$$

Resolvendo estas equações temos $a_0 = 2{,}212$, $a_1 = 1{,}299$, $a_2 = -0{,}033$; portanto, a equação da parábola é

$$y = 2{,}21 + 1{,}30x - 0{,}033x^2.$$

Uma comparação dos gráficos mostra que a parábola parece ajustar-se um pouco melhor. Pode obter-se uma medida de melhoria calculando-se o erro-padrão de estimativa para ambas as curvas. Isto é feito no problema 8.

6. Do problema 3, $\Sigma\,(y - y')^2 = 105$ quando x_2 é ignorado e $\Sigma\,(y - y')^2 = 52$ quando x_2 é usado; portanto, $s_e = \sqrt{\frac{105}{8}} = 3{,}62$ quando x_2 é ignorado e $s_e = \sqrt{\frac{52}{7}} = 2{,}73$ quando x_2 é usado. A variável x_2 é altamente benéfica porque reduz s_e a mais ou menos 3/4 de seu valor quando x_2 não é usado.

7. (a) $\Sigma x_1 = 47$, $\Sigma x_2 = 43$, $\Sigma y = 182$, $\Sigma x_1x_2 = 218$, $\Sigma x_1^2 = 277$, $\Sigma x_2^2 = 279$, $\Sigma x_1y = 972$, $\Sigma x_2y = 904$; portanto, as equações dos mínimos quadrados são

$$\begin{aligned} 10a_0 + 47a_1 + 43a_2 &= 182 \\ 47a_0 + 277a_1 + 218a_2 &= 972 \\ 43a_0 + 218a_1 + 279a_2 &= 904. \end{aligned}$$

A solução destas equações é $a_0 = 5{,}48$, $a_1 = 1{,}81$, $a_2 = 0{,}98$; portanto, a equação do plano de regressão é

$$y = 5{,}48 + 1{,}81x_1 + 0{,}98x_2.$$

(b)

y	11	21	23	7	13	18	30	18	21	20
y′	11,2	24,0	24,2	9,1	15,5	14,0	27,8	18,8	21,1	16,3
y − y′	−0,2	−3,0	−1,2	−2,1	−2,5	4,0	2,2	−0,8	−0,1	3,7

(c) $\Sigma\,(y-y')^2 = 56{,}32$; portanto $s_e = 2{,}84$.

(d) De (a) as equações dos mínimos quadrados são

$$10a_0 + 47a_1 = 182$$
$$47a_0 + 277a_1 = 972.$$

A solução é $a_0 = 8{,}42$ e $a_1 = 2{,}08$; portanto, a equação da reta de regressão é

$$y = 8{,}42 + 2{,}08x_1.$$

(e)

y	11	21	23	7	13	18	30	18	21	20
y'	10,5	23,0	20,9	12,6	18,8	12,6	25,1	14,7	23,0	20,9
y − y'	0,5	−2,0	2,1	−5,6	−5,8	5,4	4,9	3,3	−2,0	−0,9

Aqui $\Sigma\,(y-y')^2 = 142{,}5$; portanto, $s_e = \sqrt{142{,}5/8} = 4{,}22$. Como $s_e = 2{,}84$ quando x_2 está incluído, o valor de s_e é cerca de 2/3 tão grande quando x_2 está incluído como quando não está.

8. Para a linha reta

y	7,6	9,5	9,3	10,3	11,1	12,1	13,3	12,7	13,0	13,8	14,6	14,6	14,7
y'	8,6	9,2	9,8	10,3	10,9	11,5	12,0	12,6	13,2	13,8	14,3	14,9	15,5
y − y'	−1,0	0,3	−0,5	0,0	0,2	0,6	1,3	0,1	−0,2	0,0	0,3	−0,3	−0,8

$\Sigma\,(y-y')^2 = 4{,}30$; portanto, $s_e = \sqrt{4{,}30/11} = 0{,}63$.

Para a parábola

y	7,6	9,5	9,3	10,3	11,1	12,1	13,3	12,7	13,0	13,8	14,6	14,6	14,7
y'	7,9	8,8	9,7	10,5	11,2	11,9	12,5	13,1	13,5	13,9	14,3	14,6	14,8
y − y'	−0,3	0,7	−0,4	−0,2	−0,1	0,2	0,8	−0,4	−0,5	−0,1	0,3	0,0	−0,1

$\Sigma\,(y-y')^2 = 2{,}025$; portanto, $s_e = \sqrt{2{,}025/10} = 0{,}45$. Como o valor de s_e foi diminuído de cerca de 1/3, a parábola é um notável melhoramento sobre a reta.

CAPÍTULO 14

1. $E(I) = 300(0{,}6) + 100(0{,}3) - 50(0{,}1) = 205.$

$E(H) = 200(0{,}6) + 200(0{,}3) + 100(0{,}1) = 190.$

Escolha a licença para vender sorvete

2. $E(I) = 300\,p + 100(1-p).$

$E(H) = 200\,p + 200(1-p).$

$300\,p + 100(1-p) = 200\,p + 200(1-p)$. A solução é $p = \frac{1}{2}$.

3.

	A	B
S	40.000	20.000
F	−10.000	−12.000

$E(A) = 40.000(0{,}2) - 10.000(0{,}8) = 0.$

$E(B) = 20.000(0{,}4) - 12.000(0{,}6) = 800.$

Escolha B.

4.

	1.º de março	15 de março
Geada	−2.000	−2.000
Não geada	10.000	5.000

$E(1) = -2.000(0{,}6) + 10.000(0{,}4) = 2.800.$

$E(15) = -2.000(0{,}2) + 5.000(0{,}8) = 3.600.$

Escolha 15 de março.

5. Como ele sempre pode vender 100 jornais, lucrará (em cents) $E(100) = 100(4) = 400$. Como há uma probabilidade de 0,3 de que ele só venderá 100 e de 0,7 de que ele poderá vender, pelo menos, 200, se ele ordenar 200, seu lucro será $E(200) = 200(4)\,(0{,}7) + [100(4) + 100(6)]\,(0{,}3) = 500$. Há uma probabilidade de 0,2 de que ele possa vender 300 jornais; portanto, se ele ordenar 300 jornais seu lucro esperado será

$$E(300) = 300(4)\,(0{,}2) + [200(4) - 100(6)]\,(0{,}5) + [100(4) - 200(6)] = 100.$$

Ordenar 200.

6. (a)

Necessários \ Comprados	0	1	2	3
0	0	50	100	150
1	150	50	100	150
2	300	200	100	150
3	450	350	250	150

(b) $E(0) = 0{,}3(150) + 0{,}2(300) + 0{,}1(450) = 150$

$E(1) = 0{,}4\,(50) + 0{,}3\,(50) + 0{,}2(200) + 0{,}1(350) = 110$

$E(2) = 0{,}4(100) + 0{,}3(100) + 0{,}2(100) + 0{,}1(250) = 115$

$E(3) = 0{,}4(150) + 0{,}3(150) + 0{,}2(150) + 0{,}1(150) = 150$

E(1) é um mínimo; portanto, compre 1.

7.

Número vendido	20.000	25.000	30.000
Probabilidade	0,3	0,3	0,4

E (Empresariar) = [20.000(2) — 30.000(2)](0,3) + [25.000(2) + — (25.000)(2)](0,3) + [30.000(2)—20.000(2)](0,4) = = 2.000; portanto, faça o investimento.

8. (a) E(I) = 15.000(0,2) — 5.000(0,8) = —1.000.
E(N) = 0; portanto, não faça o investimento.

(b) Há quatro possibilidades aqui, se o conselho for aceito: PP, $\overline{PP}$, NP, $\overline{NP}$, onde a primeira letra expressa a produtividade do poço e a segunda, o conselho dado. Estes possíveis eventos e suas probabilidades são os seguintes:

	Probabilidade	Lucro, se o conselho for seguido
PP	(0,2) (0,6)	15.000
$\overline{PP}$	(0,2) (0,4)	0
NP	(0,8) (0,4)	—5.000
$\overline{NP}$	(0,8) (0,6)	0

E = 15.000(0,12) — 5.000(0,32) = 200. Como E aumentou de 0 (o melhor valor de [a]) para 200, o investidor deveria comprar o conselho.

(c) $ 200.

9. (a) E(A) = 20.000(0,4) + 10.000(0,6) = 14.000.
E(B) = 50.000(0,4) — 20.000(0,6) = 8.000.

Compre a menor quantia.

(b)

	Probabilidade	Lucros
LL	(0,4) (0,7)	50.000
LS	(0,4) (0,3)	20.000
SL	(0,6) (0,3)	—20.000
SS	(0,6) (0,7)	10.000

E = 50.000(0,28) + 20.000(0,12) — 20.000(0,18) + 10.000(0,42) = 17.600.
Sugere-se o valor $ 17.600 — $ 14.000 = $ 3.600.

10. (a) $P\{0,4|s\} = \dfrac{P\{0,4\}\,P\{s|0,4\}}{P\{0,4\}P\{s|0,4\} + P\{0,5\}P\{s|0,5\} + \{0,6\}P\{s|0,6\}} =$

$$= \frac{(0,3)\,(0,10)}{(0,3)\,(0,10) + (0,3)\,(0,10) + (0,4)\,(0,05)} = \frac{0,03}{0,08} = \frac{3}{8}.$$

$$P\{0,5|s\} = \frac{(0,3)\,(0,10)}{0,08} = \frac{3}{8},$$

$$P\{0,6|s\} = \frac{(0,4)\,(0,05)}{0,08} = \frac{2}{8}.$$

(b) $E = [20.000(2) - 30.000(2)]\,\frac{3}{8} + [25.000(2) - 25.000(2)]\,\frac{3}{8} +$

$$+ [30.000(2) - 20.000(2)]\,\frac{2}{8} = -2.500.$$

Não deveria empresariar o disco.

11. (a)

Evento	Probabilidade	Venda	Fabricação
Sucesso	0,7	200.000	500.000
Não sucesso	0,3	100.000	—300.000

$$E(S) = 200.000(0,7) + 100.000(0,3) = 170.000.$$
$$E(M) = 500.000(0,7) - 300.000(0,3) = 260.000.$$

Fazer o brinquedo.

$$\text{(b)}\quad P\{H|S\} = \frac{P\{H\}P\{S|H\}}{P\{H\}P\{S|H\} + P\{N\}P\{S|N\}} =$$

$$= \frac{(0,7)\ (0,04)}{(0,7)\ (0,04) + (0,3)\ (0,03)} = \frac{28}{37}.$$

$$P\{N|S\} = \frac{(0,3)\ (0,03)}{0,037} = \frac{9}{37}.$$

Isto fornece novo quadro.

Evento	Probabilidade posterior	Venda	Fabricação
Sucesso	0,76	200.000	500.000
Insucesso	0,24	100.000	—300.000

$$E(S) = 200.000\,\frac{28}{37} + 100.000\,\frac{9}{37} = 175.676.$$

$$E(M) = 500.000\,\frac{28}{37} - 300.000\,\frac{9}{37} = 305.405.$$

Portanto, fabrique. A vantagem do lucro é maior do que o era antes da amostragem.

1989

Impressão e acabamento
(com filmes fornecidos):
EDITORA SANTUÁRIO
Fone (0125) 36-2140
APARECIDA - SP